NBS SPECIAL PUBLICATION **569**

U.S. DEPARTMENT OF COMMERCE / National Bureau of Standards

Validation and Assessment Issues of Energy Models

NATIONAL BUREAU OF STANDARDS

The National Bureau of Standards[1] was established by an act of Congress on March 3, 1901. The Bureau's overall goal is to strengthen and advance the Nation's science and technology and facilitate their effective application for public benefit. To this end, the Bureau conducts research and provides: (1) a basis for the Nation's physical measurement system, (2) scientific and technological services for industry and government, (3) a technical basis for equity in trade, and (4) technical services to promote public safety. The Bureau's technical work is performed by the National Measurement Laboratory, the National Engineering Laboratory, and the Institute for Computer Sciences and Technology.

THE NATIONAL MEASUREMENT LABORATORY provides the national system of physical and chemical and materials measurement; coordinates the system with measurement systems of other nations and furnishes essential services leading to accurate and uniform physical and chemical measurement throughout the Nation's scientific community, industry, and commerce; conducts materials research leading to improved methods of measurement, standards, and data on the properties of materials needed by industry, commerce, educational institutions, and Government; provides advisory and research services to other Government agencies; develops, produces, and distributes Standard Reference Materials; and provides calibration services. The Laboratory consists of the following centers:

Absolute Physical Quantities[2] — Radiation Research — Thermodynamics and Molecular Science — Analytical Chemistry — Materials Science.

THE NATIONAL ENGINEERING LABORATORY provides technology and technical services to the public and private sectors to address national needs and to solve national problems; conducts research in engineering and applied science in support of these efforts; builds and maintains competence in the necessary disciplines required to carry out this research and technical service; develops engineering data and measurement capabilities; provides engineering measurement traceability services; develops test methods and proposes engineering standards and code changes; develops and proposes new engineering practices; and develops and improves mechanisms to transfer results of its research to the ultimate user. The Laboratory consists of the following centers:

Applied Mathematics — Electronics and Electrical Engineering[2] — Mechanical Engineering and Process Technology[2] — Building Technology — Fire Research — Consumer Product Technology — Field Methods.

THE INSTITUTE FOR COMPUTER SCIENCES AND TECHNOLOGY conducts research and provides scientific and technical services to aid Federal agencies in the selection, acquisition, application, and use of computer technology to improve effectiveness and economy in Government operations in accordance with Public Law 89-306 (40 U.S.C. 759), relevant Executive Orders, and other directives; carries out this mission by managing the Federal Information Processing Standards Program, developing Federal ADP standards guidelines, and managing Federal participation in ADP voluntary standardization activities; provides scientific and technological advisory services and assistance to Federal agencies; and provides the technical foundation for computer-related policies of the Federal Government. The Institute consists of the following centers:

Programming Science and Technology — Computer Systems Engineering.

Validation and Assessment
Issues of Energy Models

Proceedings of a Workshop
Held at the National Bureau of Standards
Gaithersburg, Maryland

January 10-11, 1979

Edited by:

Saul I. Gass

National Engineering Laboratory
National Bureau of Standards
Washington, D.C. 20234

ABSTRACT

The Workshop on Validation and Assessment Issues of Energy Models, held
at the National Bureau of Standards, Gaithersburg, Maryland (January 10 -
11, 1979), was funded by the Energy Information Administration of the
Department of Energy (DOE), Washington, D. C. Organized by the Bureau's
Operations Research Division, the Workshop was designed to be a forum in
which the theoretical and applied state-of-the-art of validation and
assessment, with emphasis on energy models, could be presented and
discussed. Speakers addressed the following areas: DOE's activities in
assessment and validation, taxonomy and structure of assessment and
validation, the relationship between model assessment and policy research,
the Electrical Power Research Institute's Energy Modeling Forum and projects,
independent third-party model assessment, the Texas National Energy Modeling
Project, management and improvement of the modeling process, complexity of
model evaluation, definitions and structure of model assessment approaches,
model access and documentation, assessment of specific models by the M.I.T.
Energy Laboratory and other groups, energy and econometric models, and
sensitivity analysis. This volume documents the Proceedings (papers and
discussion) of the Workshop.

CONTENTS

INTRODUCTORY REMARKS

Lincoln E. Moses
Administrator
Energy Information Administration
Department of Energy
Washington, DC 20461

The Energy Information Administration (EIA) has a keen interest in the
subject matter of this workshop and a strong reason to be one of its
sponsors. These reasons come under several different rubrics. First,
there are requirements in the law for EIA to validate its models and in-
crease access to them by interested parties in the public. Second, we
hope that from this conference we will gain many clues and indications
as to how to improve the quality of some of our energy models. Third,
specifically to advance the abilities to assess and validate models will
respond to needs, both inside and outside of our organization. Fourth,
we trust that the thought that appears in this meeting will advance not
only the techniques of assessment and validation but will reach far to-
ward deeper understanding of and ability to improve modeling itself.
And, finally, we dare hope that a by-product of the larger technical
understandings to be looked for will increase our ability to take a grip
on one task which EIA regards as central to its work; that task is to
"give useful indications of the uncertainty of each forecast."

Thus, the Energy Information Administration is pleased to participate
in the organization of the conference and in the publication of the results
of the conference and looks forward to benefits from this round, and the
possibility of further participation in future such rounds.

WELCOME

A. J. Goldman
Chief
Operations Research Division
Center for Applied Mathematics
National Bureau of Standards

Good morning. The unique capabilities of mathematical modeling, as an aid to decisions in vital public-sector areas like energy policy, are accompanied by some unique headaches. By the time one has dealt with the headaches of model design and implementation, one rarely--if I may mix my anatomical metaphors--has much stomach left for facing fully the remaining headaches of validation and assessment. These evaluative steps pose perplexing questions both conceptual and practical, questions that have forced their way to the forefront of our concerns here at the Center for Applied Mathematics of NBS.

We are therefore most grateful: to you, for coming together to participate in this most enticing program; to Professor Gass, for his hard and thoughtful work in bringing it about; and to the Department of Energy's Energy Information Administration, for proposing and supporting it. I join you in looking forward to an intellectually exciting and significant experience.

MODEL ASSESSMENT AND VALIDATION:
ISSUES, STRUCTURE AND ENERGY INFORMATION ADMINISTRATION
PROGRAM GOALS

George M. Lady [1]

INTRODUCTION

In an appendix to the recently released first volume of the
Energy Information Administration's (EIA) Annual Report to
Congress, [28], over sixty models are identified and briefly
described which can be used to project and analyze energy
production, consumption, prices and associated impacts.
Ostensibly these models represent a substantial capability in
support of the energy policy analysis process. However, generally
common features of the models are that they are large, detailed,
and in their operational form, resident on a computer. Precisely
how the results of using such models are to be interpreted and
communicated to the ultimate decisionmaker is in dispute. Serious
questions, for example, can be raised as to the actual influence
of the analysis systems developed so far. The need for developing
better procedures in these areas is reflected in the highly
successful program and attendance of this National Bureau of
Standards Workshop on Validation and Assessment Issues of Energy
Models. [2]

Questions concerning the usefulness of large computer models are
not new, and evaluations of the problems at issue are available.
The literature on model evaluation is growing, both in terms of
the generic problem and for energy models in particular. Still,
model assessments in practical terms have not been consistently
attempted until now, except perhaps in the area of military
operations research. [3] Inspection of the topics considered at
this Workshop reveals substantial differences in the opinions of
the participants as to the status of model assessment activities.
Consider that at various points during the program there are:

-   reports on the outcome of model assessment projects;

-   expressions of concern over who, generically, should
    be involved in such projects and what their roles
    should be;

-   presentations of rigorous procedures for achieving
    assessment goals;

-   a questioning of whether or not assessment goals are
    well understood; and

-   a contention that there may not be meaningful model
    assessment goals. [4]

These points of view span the spectrum between the idea that assessment projects are well specified and underway to the idea that assessment projects are not well understood and perhaps are not really feasible. Such conceptual dispersion is more indicative of the developmental status of assessment than of absolute discord.

The literature reflects the immature nature of assessment practices by failing to display a uniform terminology. Many terms are used, sometimes with a term taking on more than one meaning. Commonly encountered language includes:

- evaluation,
- assessment,
- in-depth assessment,
- verification,
- validation,
- certification,
- ventilation,
- credibility,
- documentation,
- access, and
- portability.

Some of these, "assessment" for example, remain general and intuitive while others, "verification" as an example, are converging to jargon. A uniform terminology will be beneficial and should, therefore, receive high priority.

This paper attempts to submit the currently understood model assessment issues and activities to a simple taxonomy. Having done this, the model assessment program at EIA is briefly described and related to the taxonomy. Generally, of the many concerns at stake in a "model assessment," the EIA program has reasonably complete coverage of the assessment topics so far developed in the literature. For the time being the binding constraints on model assessment appear to be the time and other resource limitations associated with accomplishing well understood and commonly agreed to tasks, rather than confusion over what should be done and who should do it.

In the next section the taxonomy is developed. Then potentially unresolved issues are discussed. Next, the EIA program is briefly described and related to the taxonomy of assessment topics. Finally, some conclusions are offered.

SOME GENERIC STRUCTURE

The subject matter of model assessment is comprised of character-istics of what the "model" represents as a resource to the analysis

process. Gass, [7], and a forthcoming U.S. General Accounting
Office report, [30], are good examples of careful efforts to
enumerate the characteristics that a model assessment should
address. Inspection of this and other literature suggests that
the many characteristics can be organized into four categories:

(1) The Model Itself

 This category identifies the mathematical and other
 logical statements that explicitly specify what the
 model does. Since the models at issue are generally
 resident on a computer, this category has two
 immediate subdivisions:

 - the intended model; and
 - the model actually resident on the computer.

 The term verification appears to be accepted as
 jargon for activities designed to determine what
 a model is and how it compares to the computer
 representation.

(2) Model Performance

 This category identifies a broad range of issues
 that can be recognized when examining specific
 assessment proposals, such as: theoretical
 sufficiency, accuracy and completeness of under-
 lying data, propriety of any estimation procedures
 utilized, and model "sensitivity." These are
 ultimately meaningful, however, only if they assist
 in interpreting model results. Exactly what concepts
 and associated measures apply to model performance
 is conceivably the major research topic in the effort
 to specify model assessment procedures. The term
 validation appears to refer to activities designed
 to determine how a model performs. In particular,
 validity refers to the issue of the degree of "fit"
 between the model and "reality." I would take this
 to mean that model validation is essentially
 accomplished through a process that somehow measures
 the "accuracy" of model results. Validation in this
 sense is thus confounded by the conceptual problem:
 models usually organize the phenomena at issue
 partially, rather than totally (i.e., model predictions
 are based upon assumptions which have their own pre-
 dictive quality). There is also the practical problem
 that many models predict far ahead in time, making it
 difficult to measure performance based upon experience
 with the model. A rigorous and accepted tradition of
 model validation procedures, therefore, is yet to be
 established.

(3) Model Uses

In the literature there appears a blur between the
idea of how a model performs, which is objective in
principle, and the appropriateness of the model in
a given use, which necessarily involves extra-model
considerations, including concepts difficult to
quantify, such as the goals (i.e., objective function)
of the model user. Following the organizing logic
presented here, a "valid" model might refer to a
model about which the "degree of fit to reality"
is known rather than is close. A level of accuracy
sufficient for some uses need not be sufficient for
all. The distinction I am drawing argues for a clear
division between the characteristics of model results
and the usefulness of model results with given
characteristics. Perhaps model validity should refer
to the first notion, while (to urge more jargon) such
as "model credibility" could refer to the second.

(4) Model Implementation (or perhaps Logistics)

This category refers to all of the attributes of the
actual use of the model. These include the various
costs and difficulties of running the model on a
computer as well as whatever staff and other resources
are necessary to maintain and make the model available.
Presumably, the ultimate issue to be resolved by a
model assessment is whether or not a particular model
is the cost-effective alternative for some (set of)
purpose(s). Accordingly, the cost of the model,
including development costs, are the appropriate
concern of an assessment.

I have found that these four categories well organize and orient
a substantial proportion of the discussions on model assessment
I have encountered. There are a number of other (perhaps meta-)
assessment issues such as: who does the assessing, what form the
results of the assessment are to take, and to whom should the
assessment results be communicated. I will note several (I
believe unresolved) issues such as these in the next section
and then go on to relate EIA model assessment activities to
the four categories described above.

SOME (POTENTIALLY UNRESOLVED) ASSESSMENT ISSUES

The Model vs Model Results

Virtually all of the discussions on assessment (including this one so far) focus upon the "model." Indeed an assessment issue arises, that of the "moving target," due to the fact that precisely what comprises the "model" can be sufficiently dynamic in the normal course of maintenance and use that any given version becomes outdated over the period of time necessary to conduct an assessment. This raises the problem of exactly which version of a model should be the object of an assessment. Aggravating this problem is the fact that the assessment process itself can stimulate model changes. Insofar as deficiencies or potential model improvements are determined during a model assessment, the modeler or model sponsor would presumably want to immediately undertake the appropriate model amendments. A persistence on the part of the assessors in calling attention to model deficiencies no longer true of a model's current version could easily upset the sometimes delicate politics among the modelers, model sponsors, model users, and model assessors.

It is my insight that a partial way around this problem is to redirect the focus of the model assessment (if possible) towards some prominent use of the model. For example, EIA has prepared comprehensive annual reports that present the results of implementing most of EIA's models. Generally, model development and updating activities key on this reporting cycle, and it is proposed that a model documentation series be developed which also conforms to the schedule associated with the reporting cycle. As a result, the particular model versions utilized for these reports would be natural choices for the versions to be considered by model assessments. The strategy of focusing assessments on model results is not only a solution to the "moving target" problem, but a constructive solution, since a prominent release of model results would necessarily be the precise stimulus for many questions the model assessment activities would be attempting to answer.

The Role of the Modeler

There is a growing opinion that the modeler must necessarily be intensively involved in the model assessment process. Since the utility of a model is absolutely constrained by the model user's understanding of, and confidence in, the model's results, standard advice has become the urging of better "communications" between model builders and model users. [5] While not daring to dispute this advice, it is my belief that establishing the "communication" at issue is a complex issue. In particular, U. S. GAO evaluation guidelines just released for review note that the nature and circumstances of modelers and model users are such that they may often be inherently segregated from one another. [6]

Indeed, the assessment process stands as an intermediary to which both can relate. I observe that the series of annual reports on energy system forecasts and analyses, now reported in the EIA Administrator's <u>Annual Report to Congress</u> have consumed the overwhelming proportion of resources devoted to modeling by the EIA analysis group and its predecessors. <u>7/</u> I must, therefore, believe that a commensurate proportion of the intended and actual use of the models involved was to be realized by the recipients of the reports. Generally, the contact of such users with modelers can only be through the outcome of the assessment process or its prerequisites (i.e., various forms of model documentation). Hence, the enhanced communication called for must be indirect and institutionalized. A fundamental burden of the model assessment process and the discipline it brings to modeling is to specify and produce materials which illuminate model results independent of the modeler (who is in essence not available) and even independent of the use of the model results (which cannot always be known in advance).

Model Certification

The goal of model assessment is <u>not</u> to certify or reject the model. There is a persistent concern that model users hope that the growing efforts in assessment will lead to clearer statements as to which models are certified as usable and which are not. I have already argued that an inherently inaccurate model might be termed "valid" so long as the precise nature of the inaccuracy is understood. It follows that a model, which is not judged suitable in one use, might be acceptable in another. <u>Any</u> alternative procedure for getting the results given by a particular model is also a model, however implicit to an individual's judgment. A goal of model assessment is, therefore, to guide the selection of a model for a particular purpose, not to certify a given model in an absolute sense.

EIA PROGRAM IN MODEL ASSESSMENT

There are three broad classes of activities within the Office of Applied Analysis, EIA, which collectively respond to the "problem" of assessing the quality and usefulness of energy models. These are:

- documentation,

- "assessment" projects, and

- "access" projects.

Each is described briefly below.

Documentation

The essential prerequisite to model assessment are those written materials which describe the model, its rationale, uses and other pertinent attributes. The documentation requirements for EIA models are briefly described by the following. [8/]

For all systems utilized in energy analysis it is the policy of the Office of Applied Analysis to maintain current and detailed descriptions of the systems, their rationale and range of applicability. All contractural efforts for the development or enhancement of Applied Analysis models or other analysis procedures are required to include as a final deliverable one or more separate documents which describe the four attributes of the model or procurement at issue as summarized below independent of any other provisions of the procurement. The contractor will prepare the documentation as separate report(s) distinct from other reports called for by the contract.

Methodology Description

This constitutes a detailed description of a model's rationale, precedent for the model in the literature, and comparison to other similar models or approaches. This level of documentation details the capabilities of the model as well as its assumptions and limitations. The basic purpose of this documentation is to explain why the model structure chosen was selected and to communicate how the model compares to and was chosen over alternatives.

Model Description

A statement of the equations and other procedures which constitute the formal model structure, a description of the data and other information utilized in developing the model structure, statistical characteristics of estimated portions of the model and any other information necessary to an understanding of what the model is and how results derived from the model are obtained.

Guide to Model Applications

A non-technical description of how to use a model for analysis or forecasting, how to specify alternative input assumptions and data, and how to interpret model output. The purpose of this documentation category is to communicate the range of issues the model is designed to address and the limitations of the model. The intended audience are those who would use model results.

<u>User's Guide</u>

This constitutes a detailed description of a model's
operating procedures including names and locations of
input files and computer programs, naming conventions,
and required job control statements. These documents
are intended for the use of EIA staff who actually
operate the model on the computer and should enable
an informed staff member to make model runs and label
his input files and output files, so that subsequent
users will be able to properly identify the files. An
annotated listing of the computer program should be an
appendix to the operating documentation. This
documentation category will require frequent revision
to be kept current.

In addition to these four categories of documentation, a fifth
"model summary" is also required which not only describes the
model, but also identifies associated bibliographic material,
responsible staff members, and summary computer system related
characteristics of the model's computer implementation. Taken
together, these documents are indispensable to model assessment.
Since the documents are to be reviewed before acceptance, it is
conceivable that the review would be so structured as to be an
<u>audit</u> of the documentation, particularly for the "methodology
description" and "model description" categories. As a result,
the process by which a model is documented might, in a substantial
measure, also constitute its <u>verification</u> in the sense defined
above.

<u>Assessment Projects</u>

As noted, precisely what is to comprise a model assessment is
still being debated. The first round of EIA sponsored assessments
(still underway) are, therefore, tasked with the dual responsibilities
of considering what, generically, is to comprise an assessment as
well as undertaking the assessment of a particular model(s).
The Scope of Work for such projects follows. Note that the term
"validation" is used in a somewhat more general sense than that
discussed above.

<center>SCOPE OF WORK</center>

The following tasks will be undertaken by _____ to specify
and apply the validation procedures.

<u>Task 1:</u> Existing documentation of the _____ analysis
systems will be examined and project personnel will establish
operating versions of the systems for project use.

<center>12</center>

Task 2: Operating and conceptual documentation will be evaluated and deficiencies identified. For the purposes of this project a documentation deficiency refers to any and all aspects of model documentation which are not available, but necessary to perform the other tasks of this project. To the extent to which documentation deficiencies exist, such remedies as necessary to support this validation project will be undertaken. It is recognized that the extent of documentation deficiencies, if any, is not now known. As a result, the remaining tasks of this project are contingent upon the successful completion of this task. The resources allocated to this task in the performance of the project, including the project schedule for this task, may differ from those specified here due to the actual extent of documentation deficiencies.

Task 3: Systems attributes will be evaluated to include:

Task 3.1: Completeness and accuracy of underlying data. This subtask calls for a finding of the sufficiency of the underlying data based upon existing documentation of the data; it does not call for an independent audit of any of the data at issue.

Task 3.2: Conceptual sufficiency of system specification. This subtask calls for a finding as to the completeness of the set of concepts or variables included in the system and the completeness of the set of interrelationships among the variables accounted for by the model. Particular emphasis is placed upon the identification of alternative specifications and the rationale for the particular specification chosen.

Task 3.3: Appropriateness of operating representation. This subtask calls for a finding as to the adequacy of the particular mathematical forms adopted for the model. Particular emphasis is placed on the functional or algorithmic forms employed for determining variable values, alternatives to such forms and the rationale for the particular forms chosen as well as those rejected.

Task 3.4: Appropriateness of embodied estimation methodologies. This subtask calls for a finding as to the adequacy of the statistical or other procedures utilized to derive the parameter values embodied in the model's mathematical representation. Particular emphasis is placed upon alternative estimation procedures and the rationale for the procedures selected for the model.

Task 3.5: System sensitivity and stability.
For each of the areas of model attribute identified under
specification, representation, and estimation this subtask
calls for a determination of the sensitivity or other
quality of model result associated with the particular
choices which make up the model itself compared to the
alternative choices not made. Particular emphasis is
placed upon a finding of the strengths and weaknesses
of the choices made compared to their alternatives.

Task 3.6: System performance compared to known outcomes.
This subtask calls for the identification of how modeling
results can be verified by comparison to known outcomes
and how the results of that comparison can be utilized in
preparing measurements or other indications of confidence
in model results. If possible, such comparisons will be
attempted. At a minimum a methodology for making and
procedure for using such comparisons will be developed.

Task 3.7: Computer related system characteristics.

Task 3.8: Any other system element or attribute which
           significantly influences the confidence in
           system results.

Task 4: The results of the evaluation will be consolidated
and a report on the system strengths and weaknesses prepared.

Task 5: A specification of alternative concepts of
"confidence" in system results will be prepared.

Task 6: A determination will be made of the relationship
between the outcome of the various system attribute evalua-
tions and the concepts of confidence. To the extent possible
a rigorous statement of this relationship will be achieved.

Task 7: A summary concept of system result confidence will
be developed to include the specification of the evaluation
activities necessary to support the determination of system
result confidence.

Task 8: An end of year report will be prepared on standards
and procedures for determining system confidence.

Inspection of the individual tasks and subtasks reveals that
Tasks 1 and 2 concern an evaluation of model documentation which,
after any remedial problems are corrected, accomplishes a model
verification. The standards for such, while stated only in
principle, are practical: the state of knowledge about what a
model "is" (including its computer implementation) is sufficient
if it enables whatever other analytical activities are required
in a model assessment.

The remaining tasks concern an assessment of model _performance_ and its relationship to the logical, mathematical, statistical, and other relevant model characteristics. The evident hope of the author of the task descriptions (myself) is that some form of rigorous measure of "confidence" can be specified based upon a comparison of model results to corresponding actual outcomes. For models that project far into the future precisely how, if at all, such confidence measures can be specified and made is still a research topic. If the verification were accomplished as an integral part of model development and documentation, then the work otherwise called for is model _validation_ in the sense specified above.

Access Projects

Section 113 of the Energy Conservation and Production Act (P.L. 94-385) passed in August 1976 calls for "access" to the Project Independence Evaluation System (PIES) "to representatives of committees of the Congress in an expeditious manner" and otherwise to make PIES (now termed "MEFS," Midterm Energy Forecasting System) available to the public under "reasonable terms and conditions" to include the charge of "a fair and reasonable fee...." Issues of model "access" in general were stimulated by this section of the ECPA.

In considering alternative programmatic responses to the access requirement, the spectrum of choices reduces to something like the following:

    (1)   undertake analysis projects by EIA staff at the request of others;

    (2)   establish procedures for others to implement EIA models interactively on the DOE computer; and

    (3)   establish procedures for versions of EIA models to be "transported" to another's computer system.[9]

The first of these is the current means of satisfying the legal requirement for "access." The second, to allow interactive access to models on the DOE computer by others is prohibitively expensive and is not currently being attempted. The last, the issue of model portability, is under careful scrutiny.

Independent of legal stimulus, model portability is highly desirable from the standpoint of adequate scientific method; in particular, successful model transfers ensure (or at least conclusively address) the _reproducibility_ of model results. Looking back to the idea of enhancing the "communications" between modelers and model users, such is clearly accomplished in large measure if _any_ third party can in principle duplicate all tasks appropriate to the verification and validation of a particular model version.

Model portability raises many software and hardware problems which are neither trivial nor currently viewed as conclusively solved. The current EIA program strategy is to investigate the feasibility of utilizing the National Energy Software Center (NESC) at the Argonne National Laboratory. The basic idea is that the NESC will inventory EIA models. The "portability" of the models will be certified by NESC in that prior to acceptance into their inventory NESC will successfully run the model under those circumstances specified as establishing an operating model version. The transfer of a model version to a third party will be made entirely through the NESC, independent of EIA, under the cost and other terms specified by their subscription practices. Obviously, the documentation and computer implementation standards sufficient to enable model portability in this sense are both severe and comprehensive. In particular, a practical definition of a portable model would be a model and its documentation and computer implementation necessary for a third party to undertake all analysis tasks sufficient to verify and validate the model. The budget and other aspects of the essential feasibility of this ambitious goal are not yet entirely understood.

SUMMARY AND CONCLUSIONS

The focus of this paper has been upon "how a model in use will have been assessed." The model development process has a number of stages at which assessment or evaluation of the modeling enterprise is appropriate and desirable.[10]/ Further, there are many extraordinary technical issues involved in the actual conduct of verification and validation activities. [11]/ The emphasis of this paper was not meant to minimize these.

As far as I know no energy model has been submitted to all the various verification and validation exercises identified above. As a result, credibility, an assessment of the sufficiency of a model's performance in a given use, has not yet been addressed. The EIA models now used are at the state-of-the-art. At issue in the assessment process are their objective characteristics.

The ultimate question to be addressed by the EIA program is whether or not, when the program reaches its goals, the credibility and usefulness of EIA models will have been established. A firm test of the EIA program's intentions is in the future, perhaps a year or more until any model has been assessed consistent with all the verification and validation goals discussed above. I can report that documentation deficiencies have been sufficiently extreme such that, at this point, about six to nine months since the effective start of the EIA program, no other assessment topic (except documentation) has yet been approachable. As a practical matter, documentation issues may absorb most of the model assessment resources for some time to come.

seems reasonable to expect that the EIA program, if success-
, will be a resource to model users. Yet, however successful
 current program is in this regard, the program goals will
arly have the modeler understanding the models and the model-
 results. This is a necessary first step - and a step not yet
en.

DISCUSSION

. Goldman (NBS): I would just like to comment on the term "port-
ility." These thoughts are suggested by the fact that at the
tional Bureau of Standards we have a Center for Consumer Product
chnology. One of the topics it was asked to study was the
propriateness of advertised claims that certain TV sets were
rtable. You get into some interesting questions, but clearly
ything is portable if you have a C5A to transport it around.

, it seems to me that one of the issues we will want to discuss
 a definition of portability is what resources are going to be
sumed.

should say that the study ended up in trying to establish for the
erage human carrier what were appropriate parameters of size,
ight, and awkwardness of shape that limited the concept of port-
ility. I don't know whether it specifically took up the existence
nonexistence of particular kinds of handles that are well-chosen
 facilitate moving something around. I suggest that possibly, in
scussions of portability of models, there may be specific questions
 be raised that are analogous to all of these points.

. Lady: No comment because we are just trying to find out what the
oblems are. So, I don't know what response I could have.

. Mayer (Princeton U.): I am just interested in your final comments
out advocates who you said wouldn't accept the model even if it is
lidated.

. Lady: I don't know whether they would accept it or just wouldn't
ke it.

. Mayer: Well, when you say they wouldn't like it, do you feel
ey are wrong? Do you feel that if a model is validated with a
oper scientific tool and had all sorts of certification then people
volved in the political process of planning our energy future ought
 like the model?

17

Dr. Lady: I don't think models could be valid themselves. No, they do not have to like the answer just because we have gone through whatever we have gone through to let it be understood what the answer represents. But the answer does represent what it does represent. If they have an alternative consequence, it seems to me that they have the very same responsibility.

So, it is not sufficient just not to like it. You have to inform about your consequence to the same degree.

Dr. Holloway(Texas Energy Advisory Council): I was wondering, on the matter of confusion about understanding terms and so, whether you thought of asking various disciplines to do some work on common understanding? It seems to me that some of the differences in approach are partly based on disciplinary orientations. If you were to ask the disciplines to be found in operations research, engineering, the social sciences, and perhaps the trade associations to pay some attention to what these terms mean coming out of that disciplinary background, it might focus some attention on the differences in understanding.

Dr. Lady: Well, that is a good idea. I am not sure how to do that. I am just hoping it gets done.

NOTES

1/ The author is the Director of the Office of Analysis Oversight
and Access, Office of Applied Analysis, Energy Information
Administration, Department of Energy. The author has benefited
from discussions with many, notably Harvey Greenberg, William
Hogan, David Nissen, and David Wood. Of course the contents
of the paper are entirely the author's responsibility and do
not necessarily reflect the opinions of others, including the
U.S. Department of Energy.

2/ Greenberger, et al, [11], note that models may not be as
intensively used as modelers suppose. Hogan, [12], feels
that modelers may expect too much.

3/ The U.S. General Accounting Office has been reviewing military
models "for a number of years," [30], p.1; this recent GAO
exposure draft and the pace setting work by Gass, [7] well
structure the assessment problem. Useful bibliographies are
given in each. [31] and [32] are examples of assessment
efforts.

4/ Assessment project reports: [1], [3], [6], [8], [13], [15],
[18], [23] and [25]; concern over participants and roles: [10],
[20], [22], [33] and [34]; rigorous assessment procedures: [9],
[16] and [24]; understanding assessment goals: [2], [17] and
[21]; feasibility of meaningful assessment goals: [14].

5/ See [12], p. 2.

6/ See [30], p. 7

7/ By these reports I mean [4], [5] and [29].

8/ The following is now standard for all EIA model development
contracts. This statement and related discussion is from an
internal DOE memorandum, C.R. Glassey (Assistant Administrator
for Applied Analysis) to Applied Analysis Senior Staff: Interim
Model Documentation Standards and Procedures, February 27, 1979
(with attachments).

9/ Shaw directed a careful consideration of the access issue.
For this Workshop's program see [25] including citations.

10/ Wood, [34], identifies a number of tempos in model development
for assessment while [19] describes many stages of model
implementation. Freedman, [6], reports on the assessment of
an undeveloped model.

11/ See especially Greenberg, [9].

REFERENCES

[1] Baughman, M.L., "Reflections on the Model Assessment Process:
     A Modeler's Perspective," Workshop Proceedings.

[2] Cazalet, E., "Energy Modeling Methods and Related Validation
     Issues," Workshop Proceedings.

[3] Cherry, B.H., "Electric Load Forecasting: Probing the Issues
     with Models," Workshop Proceedings.

[4] Federal Energy Administration, National Energy Outlook,
     Washington, D.C., February 1976.

[5] Federal Energy Administration, Project Independence Report,
     Washington, D.C., November 1974.

[6] Freedman, D., "Assessment of the READ Model," Workshop
     Proceedings.

[7] Gass, S.I., "Evaluation of Complex Models," Computers and
     Operations Research, Vol, 4, March 1977, pp. 27-35.

[8] Goldman, N.L. and James Gruhl, "Assessing the ICF Coal and
     Electric Utilities Model," Workshop Proceedings.

[9] Greenberg, H., "A New Approach to Analyze Information Contained
     in a Model," Workshop Proceedings.

[10] Greenberger, M., "A way of Thinking About Model Analysis,"
     Workshop Proceedings.

[11] _____, M.A. Crenson and B.L. Crissey, Models in the
     Policy Process, Russell Sage Foundation, N.Y., 1976.

[12] Hogan, W.W., "Energy Modeling: Building Understanding for
     Better Use," presented at the Second Lawrence Symposium on
     Systems and Decision Sciences, Berkeley, California,
     October 3, 1978.

[13] Holloway, M., "The Texas National Energy Modeling Project:
     An Evaluation of EIA's Midrange Energy Forecasting Model,"
     Workshop Proceedings.

[14] House, P.W. and R. Ball, "Validation: A Modern Day Snipe Hunt?
     Conceptual Difficulties of Validating Models," Workshop
     Proceedings.

[15] Hudson, E.A. and D. Jorgenson, "Assessment and Selection of
     Models for Econometric Analysis," Workshop Proceedings.

[16] Kuh, E. and R.E. Welsch, "Energy and Econometric Models and Their Assessment for Policy: Some New Diagnostics Applied To Translog Energy Demand in Forecasting," Workshop Proceedings.

[17] Mayer, L., "On a Perspective for Energy Model Validation," Workshop Proceedings.

[18] Murphy, F. and H.J. Greenberg, "Validity as a Composite Measure of Goodness," Workshop Proceedings.

[19] National Bureau of Standards, "Guidelines for Documentation of Computer Programs and Automated Data Systems," FIPS 38, Washington, D.C., February 1976.

[20] Nissen, D.H., "Impacts of Assessment on the Modeling Process," Workshop Proceedings.

[21] Parikh, S.C.; W. Marcuse, T. Sparrow and D. Pilati, "Appropriate Assessment" and "Validation Issues," Workshop Proceedings.

[22] Richels, R.; and David Kresge, "Third Party Model Assessment" and "An Approach to Independent Model Assessment," Workshop Proceedings.

[23] Rubin, L., and F. Hopkins, "Validating the Hirst Residential Energy Use/Mid-Range Energy Forecasting System Interface," Workshop Proceedings.

[24] Schweppe, F. and J. Gruhl, "Systematic Sensitivity Analysis Using Describing Functions Models," Workshop Proceedings.

[25] Shaw, M., "Model Access and Documentation," Workshop Proceedings.

[26] Stauffer, Jr., C.H., "Developing, Improving and Assessing the ICF's Coal and Electric Utilities Model," Workshop Proceedings.

[27] Sweeney, J., "The Energy Modeling Forum," Workshop Proceedings.

[28] U.S. Department of Energy, Energy Information Administration Annual Report to Congress 1978, Volume I, U.S. DOE, Washington, D.C., 1979.

[29] _____, _____ Annual Report to Congress, Volume II (1977) Projections of Energy Supply and Demand and Their Impacts, U.S. DOE, Washington, D.C., April 1978.

[30] U.S. General Accounting Office, Guidelines for Model Evaluation (Exposure Draft), PAD-79-17, U.S. GAO, Washington, D.C., January 1979.

[31] _____, An Evaluation of the Use of the Transfer
     Income Model -- Trim -- To Analyze Welfore Programs, PAD-28-14
     U.S. GAO, Washington, D.C., November 25, 1977.

[32] _____, Review of the 1974 Project Independence
     Evaluation System, OPA-76-20, U.S. GAO, Washington, D.C.,
     April 21, 1976.

[33] Weyant, J., "The Energy MOdeling Forum and Model Assessments:
     Substitutes or Compliments," Workshop Proceedings.

[34] Wood, D.O., "Model Assessment and the Policy Research Process:
     Current Practice and Future Promise," Workshop Proceedings.

Model Assessment and the Policy Research Process:
Current Practice and Future Promise

David O. Wood
M.I.T. Energy Laboratory
and Sloan School of Management

Introduction

The rapid increase in the development and application of large-scale
energy policy models since the 1973-74 OPEC oil embargo is unprecedented
in the policy sciences. While other public policy areas, such as urban
planning and water resources planning, have stimulated intensive modeling
and model application efforts, energy policy modeling seems more visible
and to have stimulated both the enthusiasm and concerns of broader
constituencies. Visibility of energy policy modeling seems due both to
the pervasiveness of energy in society and the perceived urgency of
energy issues, and to active programs in government industry,
foundations, and universities to develop and apply policy models in well
publicized studies.* Such studies published in a form highlighting the
role of energy policy models focus attention on models, sometimes at the
expense of the analysis itself. The early identification of the FEA
Project Independence Evaluation System (PIES) as the "pocket pistol of
the President" gives some flavor of one type of concern about the role of
models in the policy process.

While government and quasi-government model-based policy studies
have contributed significantly to model visibility, these applications
have taken place in a broader context of scientific research and analysis
of energy production and use. Prior to the embargo the NSF Research and
Analysis for National Needs (RANN) program was sponsoring many energy
related research projects, projects which were greatly stimulated by the

---

*Examples would include the Project Independence Report [10]; the ERDA
studies of National Research and Development [37]; the National Energy
Outlook-76 report [35]; the EIA Annual Administration Reports [17]; and
the ERDA sponsored NAS study on Nuclear and Alternative Energy Systems;
the Ford Foundation Energy Policy Project [60], and the report of the
Nuclear Energy Policy Study Group [59]; Baughman and Joskow's analysis of
the future of the U.S. nuclear industry [52]; Hudson and Jorgenson's
analysis of the likely macro-economic and energy sector effects of
alternative energy tax policies [57]; and the MIT Policy Study Group
study of conditions for energy self-sufficiency [58].

focussing power of the embargo.* Subsequent to the embargo these
modeling research activities were expanded with sponsorship from the DOE
predecessor agencies, and from EPRI, various foundations, and
universities.

In addition to modeling research the policy interests of government
and industry have stimulated the formation of commercial firms oriented
toward providing model-based support for energy policy evaluation and
analysis. Although generalizations are dangerous, as a rule these firms
tended to organize and apply academic research results in modeling
efforts and studies aimed at particular client groups and/or policy
evaluation and analysis issues.**

This large investment in energy research and modeling has been based
upon, and has stimulated, confidence that policy models can make a major
contribution in energy policy evaluation and analysis. That the essence
of a policy issue is that differences in policy turn on value conflicts
between two or more constituencies in resolving factual or analytical
disputes and/or in interpretation of expected consequences of
implementing particular policies is generally recognized; but even when
value conflicts dominate, model proponents argued that systematic
analysis and presentation of the results of alternative policies helps to
make clear the nature and extent of the value conflict.

But while expectations are high, the actual success of model-based
policy evaluation and analysis has not yet been widely demonstrated and
accepted. The sources of disappointment are not readily classified, but
seem mostly related to perceived failures in the models themselves and in
the policy modeling process. Caricaturing somewhat; analysts tend to
find models unfocussed and lacking detail for the specific issues of

---

*Much of the pre-embargo modeling effort is described and/or referenced
in Macrakis [55].

**There are many examples. Firms represented at this conference which
have been particularly prominent would include ICF, Inc., Decision Focus,
Inc., Data Resources, Inc., and Dale Jorgenson Associates.

interest, and difficult to "reconfigure" in a timely manner; further, assurances about the scientific validity of energy policy models have not been satisfactory. In contrast modelers are frustrated by the elusive and changing nature of the issues as posed by policy analysts, and sometimes even suspect the rationality of the policy process. Decision makers who rely on analysts as well as their various constituencies for inputs to policy making are confused and alarmed by conflicting analyses and are led to suspect the integrity of the modeling and analysis process. Finally the various constituencies potentially affected by model-based policy analysis seem suspicious that the modeling process may be indirectly resolving disputed factual, analytical, and value conflicts in a "blackbox" environment, especially when the model results are inconsistent with their policy positions.

Important initiatives to deal with these concerns include efforts to ensure congressional oversight of government energy modeling and data development, and government and industry efforts to better organize model-based policy research and to ensure scientific review and analysis of policy models, with results communicated in a form accessible to all groups interested in energy policy research. The following remarks focus upon recent and current activities related to such policy model analysis, including a survey of the activities of government, industry, and universities relating to policy model evaluation and analysis, and some speculations about the future of these activities.

2. Recent and Current Activities Relating to Organization and Conduct of
   Energy Policy Model Analysis and Evaluation.*

   We now turn to a review of the activities of various organizations
concerned with the development and use of models in the energy policy
research process as they relate to improving policy model credibility and
utility. Most prominent are the Congress, the Department of Energy (and
its predecessor agencies), the Electric Power Research Institute (EPRI),
and the National Science Foundation (NSF). In addition the General
Accounting Office (GAO), the National Bureau of Standards (NBS), and the
Texas Energy Advisory Council (TEAC) have been significantly interested
and involved in policy model analysis activities.

   2.1 Congressional Activities

   Perhaps the most intriguing activities relating to policy model
analysis and credibility are those of the U.S. Congress. Following the
publication of the first Project Independence Report [10], the Congress,
both in hearings and legislation, expressed concern about the credibility
of available energy data and of studies and analyses using those data,
and in particular the Project Independence Evaluation System (PIES).
They feared that analysts within the government were too closely related
to the energy industry—in particular the petroleum industry—to prepare
truly independent and objective reports; that the Executive Branch was
exerting influence on the data development analysis efforts to support
particular Administration policy positions; and that the assumptions and
interpretations of particular analyses were not well grounded in
scientific knowledge, were not well documented, and could probably not be
replicated.

   These concerns led to a number of related Congressional actions.
First, the Energy Information Administration (EIA), was organized so as
to insulate the energy data and analysis functions from the policy
formation and analysis functions of the Department of Energy [1]. This

---

*Material in this and the following section is drawn primarily from a
forthcoming report to the EIA Office of Analysis Oversight and Access
from the M.I.T. Energy Laboratory.

was accomplished in part by organizing EIA as an Administration with direct reporting responsibilities to Congress.* Perhaps most significant of these reporting requirements is the Administration's Annual Report, presenting short-, medium-, and long-term analyses of energy supply, demand, and consumption independent of the policy analysis function of the DOE [17].

Second, Congressional and public access to the PIES and related models on "reasonable" terms was mandated [1]. In addition the GAO was asked to provide an assessment of the PIES [15], and the House Subcommittee on Energy and Power commissioned its own PIES evaluation [40].

Third, an independent Professional Audit Review Team (PART) was established to conduct an annual audit of EIA activities, and to report its findings directly to Congress [1].** The first PART report, published in December 1977, apparently confirms many of the Congressional concerns. Thus,

...the credibility of OEIA's [now Energy Information Administration] models has not been established because documentation, verification, and validation have been neglected. Furthermore, publications describing the current models are scarce, and procedures for public access to them are almost nonexistent. As a result, it is partically impossible for interested parties outside FEA [now part of the Department of Energy] to know whether OEIA's current models have been constructed properly and used correctly and thus whether OEIA's analytical products and forecasts can be used with confidence [53].

The report also questions EIA's procedures in distinguishing model development activities from model applications, and makes a series of recommendations including improved documentation, better control of model

---

*Among others, Hogan [41] has argued that this concern with ensuring the integrity of the energy data and analysis function has tended to greatly reduce the effectiveness of EIA, and its relevance in the policy research process.

**The PART is composed of a representative from each of six agencies including the General Accounting Office, (Chairman), the Securities and Exchange Commission, the Bureau of Labor Statistics, the Federal Trade Commission, the Bureau of Census, and the Council of Economic Advisors.

changes, validation of model structure, verification of model implementation, sensitivity testing to increase understanding of model response to changes in data inputs, and increased public participation of researchers outside FEA in professional review.

These Congressional actions are unprecedented and suggest the importance that the Congress attaches to ensuring the integrity of the energy data acquisition and analysis functions. The concerns about adequacy of documentation, public access to government based models, and credibility of analysis all indicate that the Congress is deeply concerned about the role of policy models in energy policy research and decision making.*

2.2 Energy Information Administration (EIA) Activities

The primary object of Congressional concern, the Energy Information Administration (and its predecessor agencies) has undertaken a variety of actions to address these concerns, and to develop and implement "good scientific practice" as part of their model development and application programs. Partially in response to the PART recommendations, EIA has established the Office of Analysis Oversight and Access. The mission of EIA/OAO&A has been to develop and implement procedures for internal management and control of the model development, application, and assessment process. The EIA/OAO&A has undertaken assessments of important EIA models, as well as formulating and implementing procedures to facilitate documentation and public access. These actions include promulgation of interim documentation standards [2], and developing plans to transfer the PIES system and associated data as well as other important EIA energy models, to the Argonne National Laboratory Software Center as a means of facilitating public access. More immediately, the EIA has been responsive to the requests of the Texas Energy Advisory

---

*It is beyond the scope of these remarks to pursue further Congressional activities relating to energy data, policy models, and policy model applications. Suffice it to say that no other agency in the federal statistical establishment has been subjected to such intense scrutiny, nor have their activities been so circumscribed as have EIA's.

Council (TEAC) for assistance in transferring PIES to Texas A & M University in support of an independent, in-depth assessment by TEAC.*

In addition to the various organizational initiatives described above, EIA has undertaken and/or supported a number of assessments of important agency models. The first major assessment supported by EIA (at the time FEA) was the Resources for the Future assessment of the version of PIES used in the First National Energy Outlook [3]. More recently EIA has undertaken assessments of the Regional Energy and Demographics Model (READ) [4], the National Coal Model (NCM), the Oil and Gas Submodels of the PIES, and the Electric Submodel of PIES**. Finally EIA has supported a study by Logistics Management Institute (LMI) to analyze the alternative means by which EIA can respond effectively to the public access requirements mandated by Congress [54], and is currently supporting a study at M.I.T. on procedures for internal management and control of model development, applications, and evaluation.

Thus specific Congressional concerns are being systematically addressed by EIA. It remains to be seen how the PART will evaluate these initiatives in its subsequent reports.

2.3   National Science Foundation (NSF) Activities

The NSF has supported a number of model evaluations as well as research on influential policy models which have helped shape current practice and understanding of the model assessment process. Although not specifically concerned with energy models, an oft-cited NSF-sponsored study by Fromm, Hamilton, and Hamilton [5] surveyed modelers and model sponsors on model characteristics, documentation, and actual use in supporting policy research. The questionnaire for project director obtained information on general description of model, model development,

_____

*See the paper by Holloway [33] contributed to this conference for a discussion of the TEAC project to evaluate PIES.

**These model assessments and three subsequently mentioned in this section will be described in greater detail in Section 3.

cost of development, planning factors and data, supporting facilities, documentation, model utilization, model assessment, and opinions concerning various policies to facilitate model development and application. The questionnaire for agency sponsors included questions about agency rationale for supporting the model development, cost and funding, model utilization, model assessment, and opinions concerning policies to influence model development and application. The study is often cited for its evidence regarding poor documentation of models, and relatively low utilization. While not detailed with respect to opinions about approaches to validation and verification, the survey did include questions on opinions concerning a model clearing house, standardized routines and procedures, federal standards and procedures for validating and evaluating models, and validation review boards. In general modelers and model sponsors tended to have complementary views, exhibiting some support for the idea of a model clearing house and for standardizing computer routines and algorithms, and opposing federal standards and review boards for validation, with review boards being slightly favored over standards. While the survey produced much useful information, it was not really focused on obtaining opinions on how to improve the utility of policy models, and provided no scope for allowing respondents to indicate their views as to what would constitute good practice and procedures for policy model validation and verification.

The NSF, together with the Russell Sage Foundation, also supported another important research effort, the study by Greenberger and his colleagues on models in policy research [7]. While not explicitly concerned with energy models, Greenberger et al. considered the role of modeling in policy research and, through case studies, the circumstances likely to influence the success or failure of such modeling efforts. Case studies were developed for models employing different methodologies and a detailed analysis of the New York City-Rand Institute as a policy research organization was undertaken. Based upon their analysis and case studies, Greenberger and his colleages arrived at rather harsh

conclusions concerning the present state of policy modeling, and the reasons for low credibility of such models. Thus,

> Professional standards for model building are nonexistent. The documentation of models and source data is in an unbelievably primitive state. This goes even (and sometimes especially) for models actively consulted by policy makers. Poor documentation makes it next to impossible for anyone but the modeler to reproduce the modeling results and probe the effects of changes to the model. Sometimes a model is kept proprietary by its builder for commercial reasons. The customer is allowed to see only the results, not the assumptions. [7, p. 338]

To rectify this situation, Greenberger et al. believe that a new professional activity needs to evolve.

> What we do propose, however, is the development of a new breed of researcher/pragmatist--the model analyzer--a highly skilled professional and an astute practitioner of the art and science of third-party analysis. Such an analysis would be directed toward making sensitivity studies, identifying critical points, probing questionable assumptions, tracing policy conclusions, comprehending the effects of simulated policy changes, and simplifying complex models without distorting their key behavioral characteristics. [7, p. 339]

> The model analyzers would be neither model builder nor model user, but in a middle position between the two, empathetic to both. [7, p. 339]

This proposal, the development of a professional interest in third-party assessment of policy models, is often cited as an important stimulus to the development of model assessment. However, Greenberger et al. provide little more information on how such a professional activity is likely to evolve, other than to note the problem of professional incentives. It has remained for DOE and EPRI to begin the process of stimulating model analysis and assessment research activities.

NSF has sponsored a number of model evaluation efforts, including the reviews of the first version of PIES by the MIT Policy Study Group [8] and the Battelle Memorial Institute [9], and a review by SRI of six energy/economy models [12]. Interest in the PIES evaluations was considerable, with the principals presenting testimony before the Joint Economic Committee [11]. The SRI review which develops considerable

pedagogical information relating to energy/economy modeling has also been influential.

In addition to supporting research on models in the policy process and actual model evaluations, NSF has supported two conferences relating to model validation, including the Conference on Model Formulation, Validation, and Improvement held in Vail, CO, June 14-15, 1975 [13], and the Workshop of Validation of Mathematical Models of Energy-Related Research and Development held at Texas Christian University, June 21-23, 1978 [14].

2.4 Electric Power Research Institute (EPRI) Activities

At about the same time that the Congress was legislating the PART into existence, EPRI began a series of activities relating to model assessment and analysis, and to improving understanding of the scope for energy policy models in the policy research process.

Perhaps the most significant of the early EPRI activities was sponsorship of the EPRI-Stanford Workshop for Considering a Forum for the Analysis of Energy Options Through the Use of Models [44]. The purpose of the workshop was to discuss and plan a forum which would provide a means for organizing studies involving modelers and model users in the analysis of selected energy problems. The activity, subsequently titled the Energy Modeling Forum (EMF) represents a most creative "invention" for organizing policy research. The activities of the EMF thus far include contributions to important energy issues including a study of the relationship between energy and the macroeconomy, a study of the issues surrounding the role of coal in the transition to new energy sources, and a study of models in electric utility planning.* Additional studies are currently under way on a survey and analysis of energy demand elasticities, and a study of oil and gas exploration and production models is being conducted. The EMF has spawned a similar EPRI Forum concentrating on electric utility models, and the general style of the EMF has been adopted in an experimental "Energy Policy Forum" activity

---

*See the paper by Sweeney [45] included in these proceedings for a discussion of EMF activities.

organized within the Department of Energy. The EMF studies are widely recognized as making a significant contribution in the analysis of important energy issues as well as representing a successful initiative in organizing modelers and model users in conducting policy research.

In parallel with the Forum, EPRI has also sponsored the M.I.T. Model Assessment Group, an experiment in alternative approaches to independent model assessment. Independent, or third-party, model assessment was discussed at the EPRI-Stanford Workshop, and its role in relationship to the Forum activity was summarized as follows.

> The panel described the role of third-party model analysis as a complement to Forum studies. The Forum must exploit the backroom concept of Forum operations, relying on the model developers to implement and translate the scenario specifications. The significant practical advantages of the procedure are achieved at the loss of the advantage of constructive independent investigation of model structure and operation. This activity supports the objectives of the Forum effort, but requires a different environment with intense involvement of individual analysts. The contributions of third-party assessment can be pursued independently [4, p. II-19].

As an outgrowth of this discussion, the M.I.T. Energy Laboratory organized the Model Assessment Group to undertake, with EPRI sponsorship, independent assessment of two important energy policy models, the Baughman-Joskow Regionalized Electricity Model, and the Wharton Annual Energy Model [24]. As a consequence of this study, the group identified a number of key issues in organizing and conducting policy model evaluations. EPRI has also sponsored assessments of energy models of special interest and importance to the electric power sector. These include an assessment by Charles River Associates, Inc. of models of the demand for electric energy [31], assessment of studies of coal supply [32], a review of the Brookhaven National Laboratory's model relating to electric utility R&D planning [30], and comparative assessments of natural gas supply models [39].

2.5 General Accounting Office (GAO) Activities

The GAO has both conducted model evaluations and contributed guidelines for model validation and verification. Following publication of the Project Independence Report [10] and at the request of Congress,

GAO prepared an evaluation of PIES based upon the M.I.T. and Battelle reviews and upon their own independent evaluation [15].

The GAO has proposed guidelines for model evaluation [16] intended to increase credibility and usefulness of policy models, and to promulgate good scientific practice in policy model development and application. The GAO guidelines are closely related to the efforts of Gass [34], emphasizing evaluation of model documentation; theoretical, data and operational validity; and verification of computer implementation. The guidelines provide an excellent discussion of issues in policy model evaluation and analysis.

2.6 National Bureau of Standards (NBS) Activities

The NBS has been involved in sponsoring research on approaches to improve policy model credibility, in developing standards for model documentation, in organizing conferences to facilitate communication among modelers, model analyzers, and model users, and in conducting policy model evaluations. With NBS sponsorship Gass [42] has considered the functions and structure of policy model documentation. He analyzes the policy model development process identifying thirteen distinct stages in a model's life cycle and proposes document types for each phase. Gass avoids recommending specific formats, recognizing that details will depend upon such factors as the purposes for which the model is intended, and need for model portability. Thus,

> Depending on the scope and ultimate use of the model, some of these documents can be eliminated or combined. In any event, the user/sponsor and the model developer must conclude an agreement as to the documents produced, their content, uses and audiences . . . The form of the documents can range from a few pages to detailed manuals [42, pp. 34-35].

NBS has also sponsored survey of the means by which the credibility and utility of policy models may be improved. The study, conducted by Computer Analysis Corp. (CAC), presented eighteen proposals to improve model utility for evaluation and comment by 39 respondents representing universities, government agencies, profit and nonprofit organizations with particular expertise in analysis, simulation, and economics [17].

The survey found the most support for propositions related to improving model initiation and implementation, and the least support for those relating to model management, in particular for a proposition to establish a federal center for model testing, verification, and validation.

In addition to conducting/sponsoring research on improving model credibility and developing guidelines for model documentation, NBS has also sponsored workshops, including the present one, and a workshop on the Utility and Use of Large-Scale Mathematical Models [18], and is presently conducting an assessment of the PIES oil and gas submodel to be discussed in the next section.

3. Survey of Recent and Current Energy Model Evaluation Studies

We now turn to the question of what constitutes current practice in
policy model evaluation. The following sections summarize approaches to
policy model evaluation, criteria of evaluation, and survey recent model
evaluation studies.

3.1 Approach to Policy Model Evaluation

As noted, the primary stimulus for the current interest in policy
model evaluation, beyond scientific peer review, is the need to
communicate evaluative information to nonscientific constituencies
concerned with model validity and applicability to particular policy
issues, and the need for organizations sponsoring model development to
develop their own standards and guidelines for good practice in both
policy model development and application. A secondary, although
important motivation is the fact that many policy models are being
developed by groups which may not have incentives for good scientific
practice such as are encouraged by peer review, reference publication,
and a strong scientific ethic.

Only recently has much attention been devoted to defining
alternative approaches to assessment and to how assessments can be most
efficiently organized. Perhaps most prominent in this regard have been
the efforts of the M.I.T./EPRI Model Assessment Project who identify four
increasingly detailed approaches to evaluation including,

> Review of literature,
>
> Overview assessment,
>
> Independent audit, and
>
> In-depth assessment.

The major distinction between the approaches concerns the materials used
in evaluation. A summary of the relationships between these approaches
to assessment is given in Figure 1*.

---

*This discussion of the M.I.T./EPRI project draws heavily upon the
Executive Summary and Chapter 1 of [24]. See also the paper by Kresge
included in these proceedings [23].

A review of the literature for a model, or set of similar models, focuses upon model formulation, measurement and estimation issues relating to model structure, applicability for analysis of specific policy issues, and so on. Such a review may be both descriptive and evaluative. A classic example is the review by Taylor (discussed below) of electricity demand models which compares model structure with an "ideal" structure. In its various forms, literature review and analysis is the traditional means of model analysis. Issues of approach, logic, measurement and interpretation are formulated and analyzed. Issues of actual implementation are less susceptible to analysis with this approach.

An overview assessment uses the underlying technical model documentation, especially the computer code, for a more precise analysis of the model's structure and implementation. An overview evaluation can identify a policy model's critical points, but it will only occasionally be able to pass judgment on the adequacy of the model's treatment of them. The overview report is a useful intermediate stage in the assessment process, but assessment of the model's validity and applicability generally requires the acquisition and analysis of experimental data.

An independent audit evaluates a model's behavior by analyzing data derived from experiments that are designed by the assessors but run by the modelers. An important element of the procedure is that the assessment group is "looking over the modeler's shoulder" while the experimental runs are being made. This is essential to the accurate interpretation of the results produced by the experiment. An audit report should use the experimental data together with the analytical material developed in previous stages of the evaluation process to determine the model's validity in as many key areas (critical points) as possible. Audit procedures have the advantages of being relatively quick and inexpensive. With complex models, however, there will generally be some critical points that cannot be fully evaluated through an audit.

An in-depth assessment develops experimental data through direct, hands-on operation of the model. Direct operation makes it feasible to carry out more complex tests, particularly when the tests require

FIGURE 1

APPROACHES TO ENERGY MODEL ASSESSMENT
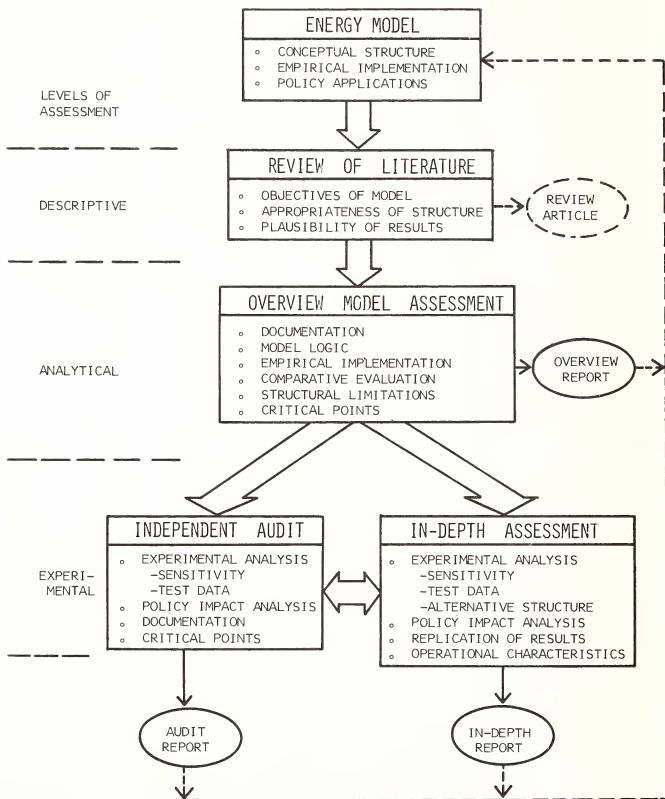
modifications in model structure rather than simple changes in model parameters and/or data. Because of the significant costs of in-depth evaluation, it is probably most efficient to conduct exploratory analysis through an independent audit before embarking on more detailed evaluation. After an in-depth evaluation has been completed, audits might subsequently be used to update the evaluation as new versions of the model are developed.

The M.I.T. group identified four separate issues regarding the organization and conduct of policy model evaluation, and proposed procedural guidelines for policy model evaluation.

Independent versus Comparative Assessments: There is a basic problem which must be confronted when evaluating a particular model as to the evaluation criteria. A model may represent the state of the art, and still be criticized for inadequacies in data and inability to deal effectively with basic research problems. Such an evaluation will provide no information establishing the strengths and limitations of the model as compared with other models potentially applicable to the same policy issues. The concern is that the potential user must be made aware of comparative strengths and weaknesses. The group concluded that this issue was a serious one, that comparative assessments should be undertaken in the future, and that the issue might become less controversial as independent evaluation became more widely practiced for developing models. In the next section several comparative model analyses oriented toward scientific audiences are noted.

Relations among Assessors, Modelers, and Sponsors: The group concluded that it is extremely important for the effectiveness of policy model evaluation that the relationships amongst assessors, modelers, and sponsors be spelled out, contractually if necessary, and that resources be provided for modeler participation in the assessment process. Modeler participation is vital to minimize the existence of misunderstanding, to correct for gaps in the documentation, and to review assessment materials. In particular, review of the evaluation and the opportunity for modelers to attach comments to evaluation reports was found to be a

vital element both in ensuring modeler cooperation, and in ensuring that the potential user is made aware of modeler disagreements with the evaluation, and with modeler/analyst perspectives.

The Moving Target Problem: Important policy models being used in policy research tend always to be developing, incorporating new research results and data, and responding to new policy research applications. This fact complicates the identification of the standard, or reference, version of the model to be evaluated. It raises questions about how model analysts should deal with significant changes to the reference version of the model. Resolving these issues requires careful planning on the part of the evaluation group, as well as flexibility and good will on the part of modelers, model analysts, and sponsors.

Assessors as Modelers: Perhaps one of the most fundamental problems in model evaluation concerns the role of the model analysts as modelers. Clearly, for evaluations to be effective and to carry weight in establishing model credibility, the assessors must be competent modelers. To the extent that assessors use the model evaluation process as a means for identifying and then pursuing research issues, however, the evaluation process is compromised. First, the modeler will lose confidence in the process and will not wish to cooperate. Secondly, the model analyst will be redirecting his activities from evaluation to research. At a minimum this represents an injustice to the assessment sponsor. Finally, the integrity of the evaluation process is seriously compromised when the assessor becomes a competitor to the modeler. Such competition is, of course, fundamental to advancement of scientific knowledge, but conflicts with objectives of communicating the results of evaluation to non-scientific audiences.

Procedural Guidelines for Model Assessment: The EPRI/MIT group suggests the following guidelines for organizing and conducting policy model evaluation.

Assessor/Modeler Relations -- A formal agreement should be reached defining the relationships between modeler and assessor with regard to:

-    resources to support modeler as well as model analysts,

- extent and nature of modeler/assessor interactions,
- confidentiality of intermediate results,
- opportunity for modeler response, and
- post-evaluation activities.

Potential Model Applications -- A wide-ranging list of potential applications of the model, incorporating suggestions from all interested parties, should be drawn up at an early stage to provide an explicit policy context for the evaluation.

Definition of a Standard Model -- A standard version of a model must be agreed upon and "locked up" prior to the start of experimental analysis. It is desirable, however, to permit changes to be made during early stages of the evaluation, particularly if the changes are to correct errors uncovered in the overview evaluation.

Assessors as Modelers -- Model analysts can and should suggest ways in which the model can be improved, but they should not themselves implement the improvements. To do so would compromise the integrity of the assessment process and would put the assessors in competition with the modelers.

3.2 Elements of Policy Model Evaluation

The elements of a policy model evaluation are much discussed and debated. The most often cited literature includes Gass [34] and Greenberger, et al. [7] both of which distinguish two fundamental aspects of model evaluation: validation and verification. Validation refers to the correspondence of the model to the underlying processes being modeled. Correspondence will include the structural features of the model, the inclusion of relevant variables—especially policy instruments and concepts of importance for the issues to be analyzed—and the predictive capability of the model. Structural evaluation is, of course, the essence of scientific analysis. Model structure is based upon the conceptual specification of the model, specification and application of the measurement process by which the model data are generated/obtained, specification and analysis of scientific hypotheses derived from theory underlying the model and to be tested via analysis of the model data, and

selection of the final model best supported by the scientific laws, principles, maintained hypotheses, and tested hypotheses which were the resources and/or results of the research process. Validation of such a research model will include replication of measurements and hypothesis testing, as well as analysis and/or counter-analysis including the variables and concepts appropriate for analysis of the policy issues for which the model is intended.

It is for this latter purpose that content validity is usually singled out from structural validity, when considering policy models. Both policy evaluation and analysis require that models reflect the appropriate policy concepts and instruments.* A policy evaluation model will be simpler than a policy analysis model in this regard in that only the policy actually implemented and being evaluated must be included. Policy analysis models are more complicated in that the policy instruments and concepts suitable for the alternative policies of potential interest and importance to the various constituencies concerned with the issue(s) of interest must be included. Further the model must be explicit concerning the resolution of "facts" and/or value judgments which are in dispute among the various constituencies. Crissey [22] has made evaluation of such model contention points a central feature of his approach to policy model analysis.

The third element of policy model validation is predictive capacity, determining if the scientific information and results included in the model are sufficient to discriminate amongst the policies being considered. If the range of scientific uncertainty spans the range of policy dispute, then the model's usefulness in policy research is very limited. Model-based studies reporting only point predictions with no information on prediction confidence limits or sensitivity analysis of predictions to changes in input data and/or structural coefficients,

---

*Greenberger, et al., distinguish policy evaluation and policy analysis as follows: "Policy evaluators organize a research effort around an existing program and ask how well it is achieving its intended objectives; policy analysts tend to organize their investigations around a set of policy objectives and they inquire whether there is any conceivable program or combination of programs that might achieve the desired ends more efficiently," [7, p. 30].

consistent with known or conjectured uncertainties in the underlying measurement processes and scientific results, may imply an unjustified precision of analysis. Analysis of predictive power is thus an important aspect of policy model analysis quite independent of the structural validity of the components of the model.

Closely related to the various dimensions of model validity is the validity of the data associated with the model. Data validation must include not only evaluation of the measurement process by which the data component of model structure is developed, but also the processes by which the data required for model applications are obtained. While data and measurement process evaluation are closely related to model evaluation, particularly evaluation of model structural and predictive capability, it is probably useful to single out this aspect of validation since it typically receives so little attention in policy modeling and research.

Crissey [22] has a similar perspective on the elements of policy model validation. He emphasizes that the credibility and utility of a policy model will depend upon its treatment of the factual, behavioral, evaluative, and structural issues in dispute. Disputed issues should be represented in the model in a manner facilitating analysis of alternative resolutions. Such issues comprise the model's contention points. According to Crissey, a contention point is said to be critical if change in its resolution significantly affects the model conclusions and is a contingency point if changing the resolution of this contention point in combination with others results in a significant change in model result [22, pp. 83-8]. This concept of model contention points provides a useful focus for structural, content, and predictive validation.

In contrast to validation, policy model verification refers to the evaluation of the actual model implementation. At issue is the correspondence of the implemented model--usually a computer program--to what the modeler intended. Verification is thus more mechanical and definitive than model and data validation. Gass [34] has suggested that policy model verification is the responsibility of the modeler, and that evaluation should be limited to review of the verification process. In

the next section, however, we will see that some of the more in-depth current policy model evaluation projects have included independent verification of implementation as an objective.

A final aspect of policy model evaluation concerns usability. This dimension of evaluation refers to both the sufficiency of documentation to support model understanding and applications, and the efficiency of the overall system. Technical documentation and materials sufficient to inform potential users of the model structure, content, and predictive characteristics, as well as to support interpretation of model-based results are essential for any policy model. The need for documentation to support independent application of the model, including user and system guides, and test problems will depend upon the model application environment. Of course, even if the intent is for the modeler to conduct all applications, there still should be evidence that application procedures have been developed, and that a reasonable applications practice is in effect.

Our discussion of the various dimensions of policy model evaluation has been impressionistic, drawing heavily upon Gass [34, 42], Greenberger, et al. [7], Crissey [22], GAO [16], the M.I.T. Energy Model Analysis Project [23, 24, 46, 47], and our review of a number of actual energy policy model evaluation efforts. We now turn to a survey of these evaluation projects classified by approach, assessor/sponsor, primary audience for which assessment is intended, and emphasis.

3.3 Survey of Model Evaluation Studies

In previous sections we have mentioned a number of policy model evaluation projects. We now turn to a nonsystematic survey and classification of these studies employing the taxonomy of validation and verification elements developed in Section 3.2. Table 1 summarizes the various energy model evaluation projects in terms of approach and emphasis.*

---

*The studies summarized are those with which I am familiar and the classification is based upon my analysis of the study reports with, in some cases, the benefit of discussion with the model analyzers. However, I am certain that some studies which should have been included are omitted due to ignorance, and the characterization of emphasis is not likely to always correspond with what the analyst intended.

TABLE 1

Classification of Energy Model Assessment Projects by
Approach, Emphasis, and Primary Audience

Assessment Project	Assessor/Sponsor	Assessment Approach[1]	Primary Audience[2]	Model Validity			Associated Data	Verification of Implementation	Usability	
				Structure	Content	Prediction			Documentation	Efficiency
Methodological Review of Six Large-Scale Energy Planning Models [12]	Brock-Nesbitt/NSF	LR	P	✓						
Methodological Review of Existing Coal Studies [32]	Gordon/EPRI	LR	A		✓		✓			
Methodological Review of Electricity Demand Models [...]	Taylor/EPRI	LR	A		✓					
Methodological Review of Energy Demand Models [25]	Hartman/Research	LR	P	✓	✓					
PIES [8]	MIT E-Lab/NSF	IA	A	✓	✓		✓			
PIES [9]	Battelle Inst./NSF	IA	A	✓	✓		✓			
PIES [15]	GAO/Congress	LR	D	✓	✓					
PIES [27]	Hausman/Research	O	P,A	✓	✓		✓			
PIES [3]	RFF/FEA	O	P,A	✓	✓		✓			
PIES [40]	Optional Analysis Co./Congress	LR	D	✓	✓					
Review of Large-Scale Energy Models [38]	CRA/EPRI	LR	P,S	✓	✓					
PIES [33]	TEAC/	I	P,A,D	✓	✓		✓			

45

TABLE 1 (continued)

Assessment Project	Assessor/ Sponsor	Assessment Approach[1]	Primary Audience[2]	Model Validity Structure	Model Validity Content	Model Validity Prediction	Associated Data	Verification of Implementation	Usability Documentation	Usability Efficiency
Comparative State-of-the-Art Assessment of Gas Supply Modeling [39]	Mathematica, Inc./EPRI	LR	P,A,S	✓						
Comparative State-of-the-Art Assessment of Oil Supply Modeling[3]	Mathematica, Inc./EPRI	LR	P,A,S	✓						
Review of Energy Models Relating to Employment and Manpower Analysis [50]	Eckstein-Hein/DoL	LR	P,A	✓	✓					
Strategic Environmental Assessment System [19]	Panel/EPA	LR	S	✓	✓					
Brookhaven Energy System Optimization Model [28]	Wood-Hausman/EPRI	LR		✓	✓					
Regional Economic and Demographic Model[3] (see also [4])	Expert Panel/EIA	LR	P,A,S	✓	✓		✓			
Comparative Assessment of Three Oil & Gas Models [21]	Pindyck/NSF	I	P	✓	✓	✓				
Wharton Annual Energy Model [24]	MIT/EPRI	IA	M,S	✓	✓					
Comparative Assessment of MacAvoy-Pindyck and TERRA Natural Gas Industry Models [29]	NERI/FEA	I	P	✓		✓				
Brookhaven Energy System Models (BESOM, DESOM) [30]	SCI/EPRI	I	P,S	✓	✓					

TABLE 1 (continued)

Assessment Project	Assessor/Sponsor	Assessment Approach[1]	Primary Audience[2]	Model Validity			Associated Data	Verification of Implementation	Usability	
				Structure	Content	Prediction			Documentation	Efficiency
Comparative Assessment of Eight Electricity Demand Models [31]	CRA/	I	P	✓	✓	✓	✓			✓
Baughman-Joskow Regionalized Electricity Model [24]	MIT/EPRI	I	P,S,A,D	✓	✓		✓	✓	✓	✓
Lawrence Livermore Laboratory Energy System Model[3]	LLL/LLL	I				✓				
Coal[2] National Energy Model [49]	Ford-Moore/EIA	I	P,A	✓	✓	✓	✓			
MacAvoy-Pindyck Natural Gas Model [22]	Crissey/Research	I	P,A,D	✓	✓	✓	✓	✓	✓	✓
ICF Coal & Electric Utility Model[3]	MIT/EPRI, EIA	I	P,A,S,D	✓	✓	✓	✓	✓	✓	✓
PIES Oil & Gas Submodel[3]	NBS/EIA	I	P,A,S,D	✓	✓	✓	✓	✓	✓	✓
PIES Electric Utility Submodel[3]	Los Alamos/EIA	I	P,A,S,D	✓	✓	✓	✓	✓	✓	✓

[1] Assessment approaches include:
LR = literature review
O = overview
IA = independent audit
I = in-depth

[2] Primary audiences include:
P = peers
A = policy analysts
S = sponsors
D = decision-makers
M = modelers

The actual classification of studies by primary audience is highly subjective, being based upon my interpretation of the study reports, and only in a few cases interactive with the model analyzers.

[3] In progress

A significant number of energy model evaluation studies based upon review of published literature and applications have been conducted. The style and focus of these studies are diverse ranging from comparative analysis of selected models against structural and content criteria derived from theoretical analysis (e.g., Taylor [20]) to reviews of model-based applications with only secondary attention to model evaluation (e.g., Gordon [32]). Further, many of these studies include several models, and so have considerable value in providing introduction to the models and interpretive material relating to their use.

Brock and Nesbitt have reviewed six large-scale energy planning models [12]. The study provides a detailed development of economic equilibrium concepts, as well as interpretive information on each of the models. The study provides less information on the appropriate application areas for each of the models.

Richard Gordon has undertaken a series of reviews of existing coal studies for EPRI which include evaluation of models, as well as model results [32]. These reviews provide an important source of information on comparing coal supply and use models. Because of their lucid presentation, these reviews are accessible to a wide audience.

Taylor [20] has conducted a review of eleven econometric models of electricity demand. This study is most interesting in that it provides a paradigm for comparative evaluation of models against criteria based upon theoretical analysis. Taylor analyzes a system characterized by regulation and decreasing block pricing. The analysis identifies concepts and variables which he argues must be considered in any model of electricity demand. Comparing the models under investigation against this standard, he finds that all are deficient respecting treatment of decreasing block pricing. He proposes an "essentially correct alternative" which he and his colleages have pursued [51]. This study is an important example of the potential importance and power of the literature review approach for comparative model evaluation

Hartman provides a methodological review of energy demand models similar in style to that of Taylor [31]. Through analysis, he identifies three characteristics which should be treated separately in any long-run

energy demand model. These include the demand for energy-related services, the demand for durable goods to combine with energy forms to satisfy energy-related service demands, and the efficiency of energy use of the durable goods. Hartman also emphasizes the importance of modeling approaches which permit the inclusion of new technologies for providing energy services. In comparing existing models with these criteria, Hartman concludes that current efforts fall short of the standard. He then sketches the theoretical and data developments required to improve existing models.

The Project Independence Evaluation System (PIES) is perhaps the most assessed and evaluated energy policy model in existence, and many of these evaluations have employed the literature review approach. The model was first evaluated following the publication of the Project Independence Report [10] by the M.I.T. Policy Study Group [8], the Battelle Memorial Institute [9], the General Accounting Office [21], and Hausman [33]. The M.I.T., Battelle, and Hausman studies are all based upon the Project Independence Report [10] and associated materials, and emphasize analysis of model structural characteristics and plausibility of results.* The GAO study, commissioned by Congress, summarizes the results of the other studies and provides perspective and interpretive material for Congressional decision makers.

Following publication of the National Energy Outlook in 1976 [52], the FEA sponsored an evaluation of the revised PIES conducted by Resources for the Future [3]. RFF organized their evaluation of PIES by convening working groups of independent experts primarily from universities. The focus of evaluation was based upon the published literature and followed the general style of the earlier Battelle and M.I.T. evaluations. However, the use of independent experts from a variety of organizations was unique to this assessment and seems to have been effective both in venting the model and model applications to a wide audience of scholars and in obtaining a comprehensive critical evaluation.

_____
*The M.I.T. study also made use of some additional runs by FEA conducted to resolve some issues relating to performance of the PIES oil and gas submodel, and so is an instance of the independent audit approach.

A second evaluation of the revised PIES model used in support of NEO-76 was commissioned by the House Subcommittee on Energy and Power and was conducted by Optimal Analysis Co. [40]. Their rather brief report focused upon the completeness of PIES to support policy decisions, and proposed new modeling efforts emphasizing dynamics of transition to new fuel forms and "crisis" (i.e., embargo) analysis.

The most recent non-EIA-sponsored assessment of the PIES modeling system (now called the Mid-range Energy Market Forecasting System) has been sponsored by the Texas Energy Advisory Council (TEAC). This evaluation, the Texas National Energy Modeling Project, involves an in-depth analysis of the component models of PIES by various research groups located in the Texas university system. The project, still under way at this time, has involved transporting the PIES system and associated data base to the computer center at Texas A&M, and replicating previous EIA analyses. In addition, the model's documentation and computer code are the raw materials for an in-depth, independent evaluation. Sensitivity experiments and the results of structural changes are being investigated. The results of this evaluation are not yet available, but promise to provide an important source of information on the validity of the PIES component models and on the accuracy of implementation. In contrast to the previous assessments of PIES, which were in the nature of literature review and overview analyses, the TEAC evaluation promises to provide a truly independent evaluation of the PIES implementation.*

Charles River Associates, Inc. has recently completed a review of 14 energy system models [38]. The focus of the review is on the applicability of each model to particular analytical and policy research issues of importance to EPRI, including fuels for generation, electricity supply, electricity use, and the environmental aspects of decisions in these three areas. The review emphasizes description and analysis of model structure and content, and in some instances provides important interpretive information not readily available from the model

─────────

*See the contribution of Holloway to this Workshop [33].

documentation. The report also includes a general discussion of selected model issues [38, Section 5].

Liliano, Limaye, and Hu [39] have conducted a comparative assessment of twelve models of natural gas supply. Three model types are distinguished including structural, econometric, and resource base-geologic, although the overlap between these various modeling approaches is acknowledged, especially in specific model descriptions and evaluations. The report provides a survey of the natural gas supply process and a general discussion of the history of modeling for this industry. Each model is described in terms of a common set of descriptors, as well as in detail, and major applications are reviewed and analyzed. The report emphasized comparative description of models, not critical evaluation, and is somewhat similar in style to the effort of Brock-Nesbitt [12]. A similar effort is presently under way by the same authors for oil supply industry models.

Eckstein and Heien have conducted a review of most of the "active" energy models with specific attention to their potential for use in employment and manpower analysis [50]. The review is based upon analysis of technical documentation and applications, and selected interaction with the modelers. The study distinguishes three periods for analysis including post-embargo shock effects, intermediate term adjustment, and long-run equilibrium. Models are classified into three groups including energy/economy, energy sector, and energy subsector models. The focus of the review is on the structural characteristics of each model's treatment of the interactions between employment and energy use. The issues and comparative model capabilities are well developed and presented in an even-handed manner.

In 1975 the Environmental Protection Agency sponsored an evaluation of the Strategic Environmental Assessment System (SEAS), a major environmental policy analysis model with important interactions with the energy sector [42]. The assessment was conducted by a panel of experts under the chairmanship of Wassily Leontief. This assessment provides the first example I have found of an assessment project for an important energy-related policy model using outside experts [25].

In 1976 Wood and Hausman conducted an assessment of the Brookhaven Energy System Optimization Model (BESOM) under the sponsorship of the Office of Technology Assessment [34]. The purpose of this assessment was to comment upon the appropriateness of BESOM in its applications in support of the first National Energy Research and Development Plan [43]. The assessment was based upon the documentation of the applications, as well as the model documentation, and concluded that the model was appropriate for the applications undertaken by ERDA, but should be extended to provide for interfactor substitution between energy and other inputs, and more direct links between the energy system and the macroeconomy.

EIA is currently sponsoring an evaluation of the Regional Energy Activity and Demographic Model (READ). The evaluation is being conducted by a panel of independent experts and consultants, primarily from universities. The assessment is based upon the available documentation for the model, as well as interaction between the evaluation panel and the modelers, who all are EIA staff members. This project is of particular interest since the READ model is in an early stage of development, with only a prototype data base and equations available at this time. Technical aspects of the assessment are discussed in the paper by Freedman [4] included in the proceedings of this workshop.

Pindyck has conducted a comparative assessment of three oil and gas supply models including the MacAvoy/Pindyck National Gas Industry Model, the FPC-Kazzoom National Gas Supply Model, and the Erickson/Spann National Gas Supply Model [27]. Pindyck reestimates the three models with a common data base and estimation procedure, and with the minimum possible structural changes necessary to put the models on a common basis. He then uses each model integrated into a complete supply/demand model to compare the regulatory policy implications of each formulation. The policy implications of the different models are quite diverse, indicating that at least among these three models no consensus exists as to how gas supplies will respond to price increases. The Pindyck study is an important example of in-depth comparative evaluation.

As part of its model evaluation activity for EPRI, the M.I.T. Energy Laboratory undertook an independent audit of the Wharton Annual Energy Model [29]. At the time of the study, only a prototype version of the model was available, and little documentation existed. The study consisted of setting up a series of computational experiments in which the expected performance of the model was hypothesized. The computational experiments then served to confirm these hypotheses, and/or to provide material for analysis of model behavior.

Neri has conducted a most interesting comparative assessment of the MacAvoy/Pindyck Natural Gas Industry Model and the American Gas Association Terra Model [35]. Neri focused on developing simulations from the two models which were normalized to the greatest extent possible respecting input data. He then analyzed the model data and structure in an effort to explain the differences in the two forecasts, and provides a detailed reconciliation of the simulations expressed in terms of differences in model data and structure. The study is an excellent example of comparison of models with differing structures, associated data, and methods for estimating structural parameters.

Systems Control Inc. (SCI) has recently completed an in-depth assessment of the Brookhaven Energy System Optimization Model (BESOM), and the Dynamic Energy System Optimization Model (DESOM) [36]. The objectives of the SCI study were to analyze the potential of these models for R&D planning in the electric utility industry, and to implement specific modifications to improve model performance. Thus the study had a dual objective of evaluation and modeling.

Charles River Associates (CRA) has conducted a study evaluating eight econometric models of electricity demand. Their approach is as follows: "Each model is replicated, reestimated on a common data set and tested for performance; forecast and backcast accuracy; parameter stability over time; robustness of parameter estimates to small changes in specification or variable measurements, consistency and plausibility of model results, and quality of model test statistics." [37, p. iii]. The study represents a major effort in putting models on a comparable basis in terms of data and estimation procedures, while preserving

original structural specifications. The study results are complicated and difficult to summarize. However, the study represents a major accomplishment in comparative model analysis.

The M.I.T. Energy Laboratory has conducted an in-depth evaluation of the Baughman/Joskow Regionalized Electricity Model [29]. The study focuses upon the model applications proposed by the modelers or likely to be considered by potential users, and attempts to evaluate the models likely success in supporting these policy analysis applications. The documentation of the evaluation is organized so as to satisfy the interests of scientists, policy analysts and decision makers. In addition to evaluating model validity and verifying implementation, the study assesses the usability of the model and includes recommendations for improving documentation and user efficiency.

Ford, Moore, and McKay have recently completed an evaluation of the COAL2 National Energy Model [49]. Their study emphasizes the application of methods to analyze the implications of uncertainties in input data and parameters upon output variables. The study does not provide much information on model validity beyond that already provided in the model documentation.

An especially interesting example of a modeler/user policy model evaluation is a project currently in progress at Lawrence Livermore Laboratory (LLL). In a separate effort LLL has transferred, adopted, and reprogrammed the Gulf-SRI Energy System Model [48]. In an effort to gain understanding of the model's predictive capacity, LLL has set up an associated data base for the period 1950-present, and is simulating the model over that period for the purpose of analyzing its predictive performance. This effort is especially noteworthy in that the model structure is based primarily upon analysis of engineering and economic data, on submodels of engineering processes, and on analysis of energy industry expansion and operating plans. Typically it has been argued that historical simulation for such models is very complicated, if not impossible, because of the difficulty in obtaining historical data, especially industry plans and expectations data, which are independent of subsequent events. The LLL effort thus represents a major undertaking,

and the results will be of great interest, both as to what is learned about model performance and what is learned about prediction analysis of engineering process/energy system models.

Crissey has conducted an evaluation of the MacAvoy/Pindyck Natural Gas Model to illustrate various concepts he has formulated regarding the usefulness and effectiveness of models in the policy research process. Crissey formulates a statement of structural features expected in a model of the natural gas industry to be used in analyzing such policies as price deregulation. Separate lists of characteristics for policy analysis capabilities and instruments are developed, and an analysis of the policy debate is undertaken to identify contention points. All this information is then used to distinguish and analyze the model's contention points. In addition, Crissey makes a separate contribution in demonstrating that the MacAvoy/Pindyck model can be greatly simplified through certain approximations and aggregations, which have little effect upon predictive behavior.

Finally, the EIA Office of Analysis, Oversight, and Access is currently sponsoring a number of model analyses including the Mid-Range Energy Market System oil and gas submodel (NBS) and electric utility submodel (Los Alamos), and the Coal and Electric Utility Model (M.I.T.).* These efforts are of special interest because they include the objective of emphasizing all of the dimensions to policy model evaluation identified in Section 3.2.

---

*See the paper by Goldman and Gruhl included in these proceedings [46].

4. The Future of Policy Model Analysis: Some Speculations

The purpose of these remarks, coming near the beginning of the workshop, has been to survey the factors giving rise to concerns about policy model credibility; to report the relevant research and organizational initiatives addressing those concerns; to summarize the elements of policy model validation and verification being applied in current evaluation studies; and to survey and briefly summarize recent and current evaluation projects. As is apparent, much activity is under way, and many of the initiatives and projects briefly mentioned here will be dealt with in greater detail in other presentations.

Having provided an introduction I now would like to offer some speculations on the future directions of policy model analysis activities. First, it seems clear that the various efforts to better organize energy policy research, and to improve the credibility of policy models are not temporary phenomena. Legislation for Congressional oversight, organizing the EIA/ OAO&A, and organization of the EPRI/Stanford EMF and the EPRI/MIT Energy Model Analysis Program all suggest a continuing commitment to these activities.

Second, I'm optimistic that the efforts of EIA, NBS, and others in developing guidelines for policy model documentation will contribute significantly to strengthening this "Achille's heel" of policy modeling. The serious efforts of model sponsors and policy modelers in evolving these guidelines into accepted "good practice" seem assured.

Third, it also seems clear that the various Forum initiatives are filling an important function, and will certainly expand in depth of analysis and significance of contribution. Reviewing the sequence of EMF studies suggests a trend toward policy model analysis and pursuit of the scientific issues raised by that analysis. I predict that these trends will result in future forum-like groups combining policy modeling and policy issue analysis, with the involvement of modelers, analysts, decision makers, and affected constituencies.

Finally, one aspect of policy model analysis that I believe will disappear is the distinction, and suggestion of competition, between the forum-like activities and policy model evaluation studies. Concern about scientific validity and applicability of energy policy models, and the response of model sponsors--such as DOE and EPRI--to the demand to "do something" may suggest to some that policy model evaluation is an alternative to better organization of the policy research process. In my opinion model evaluation is a necessary condition for more credible policy research, not a substitute. Model evaluation activities are aimed at extending traditional peer review and scientific analysis to the policy sciences. The apparent differences between peer review and current policy model evaluation efforts are due to the need to present evaluative results in a form accessible to the non-modeler group involved in the policy research and policy making process. The fundamental activity is still scientific analysis.

I believe that as the means of developing and presenting evaluative information improves and becomes accepted practice in the policy sciences, that the apparent distinctiveness of policy model evaluation will decrease. As David Nissen suggested during his presentation, policy research and policy research evaluation are inseparable parts of policy making. If policy models are to play a significant and constructive role in policy research, then the evaluative function must be satisfied in a manner consistent with both good scientific practice and the legitimate information needs of all groups involved in the policy process. This suggests to me that the development of policy model analyzer professional activity, as suggested by Greenberger, et al [7] will take place as part of the maturing of the policy sciences. The learning and incentives for such a professional activity will be based upon commitment to policy research, not just to policy model evaluation. How the professional skills and orientation will evolve is speculative; that they must is a necessity for the credibility and usefulness of model-based policy research.

REFERENCES

1. Department of Energy Organization Act, P.L. 95091, enacted August 2, 1977.

2. Energy Information Administration, "Memorandum for Applied Analysis Senior Staff," from George M. Lady, through C. Roger Glassey, subject: Interim Model Documentation Standards, December 4, 1978.

3. Hans H. Landsbert (ed.), "Review of Federal Energy Administration National Energy Outlook, 1976," a report prepared for the National Science Foundation by Resources for the Future, Washington, D.C., March 1977.

4. David Freedman, "Assessing the Regional Energy Activity and Demographic Model," (this proceedings).

5. Gary Fromm, William L. Hamilton, and Diane E. Hamilton, "Federally Supported Mathematical Models; Survey and Analysis," National Science Foundation, Washington, D.C., 1974.

6. Saul Gass, et al., "A Study for Assessing Ways to Improve the Utility of Large-scale Models," Control Analysis Corp., report prepared for the National Bureau of Standards, December, 1978.

7. Martin Greenberger, Matthew Crenson, and Brian Crissey, Models in the Policy Process: Public Decision-making in the Computer Era, Russell Sage Foundation, New York, 1976.

8. M.I.T. Energy Policy Study Group, "The FEA Project Independence Report: An Analytical Review and Evaluation," MIT Energy Laboratory Report, May, 1975.

9. Battelle Memorial Institute, "A Review of the Project Independence Report," report submitted to the Office of Energy R&D Policy, National Science Foundation, January, 1975.

10. Federal Energy Administration, Project Independence Report, Government Printing Office, Washington, D.C., November, 1974.

11. Hearings before the Joint Economic Committee, "Reappraisal of Project Independence," U.S. Senate, March 1975.

12. Horace W. Brock and Dale M. Nesbitt, "Large-scale Energy Planning Models; A Methodological Analysis," Stanford Research Institute, Menlo Park, California, May, 1977.

13. National Bureau of Economic Research, Inc., "Model Formulation, Validation, and Approvement," Proceedings of an NSF Sponsored Conference held in Vail, Colorado, June, 1975.

14. C.R. Deeter and A.A.J. Hoffman, (eds), "Validation of Mathematical Models of Energy Related Research and Development," Proceedings of an NSF Sponsored Conference held in Dallas, Texas, June, 1978.

15. General Accounting Office, "Review of the 1974 Project Independence Evaluation System," report to the Congress by the Comptroller General of The United States, 1975.

16. General Accounting Office, "Guidelines for Model Evaluation," U.S. General Accounting Office, Washington, D.C. (PAD-79-17), January, 1979.

17. Energy Information Administration, "Annual Report to Congress, 1978," (DOE/EIA-0173) in 3 volumes, 1979.

18. Saul Gass (editor), "Utility and Use of Large-scale Mathematical Models," proceedings of a workshop held at the National Bureau (NBS Special Publication 534), May, 1979.

20. Lester D. Taylor, "The Demand for Electricity: A Survey," The Bell Journal of Economics. Volume 6, number 1, (Spring), 1975, pp. 74-110.

21. Robert S. Pindyck, "The Regulatory Implication of Three Alternative Econometric Supply Models of Natural Gas," The Bell Journal of Economics and Management Science. Volume 5, number 2 (Autumn), 1974, EP 633-45.

22. Brian L. Crissey, "A Rational Framework for the Use of Computer Simulation Models in a Policy Context," unpublished PhD. dissertation, The Johns-Hopkins University, 1975.

23. David Kresge, "Third Party Model Assessment," (this proceedings).

24. MIT Model Assessment Group, "Independent Assessment of Energy Policy Models," Electric Power Research Institute, Palo Alto, California, (EA-1071), May, 1979.

25. Raymond Hartman, "Frontiers in Energy Demand Modeling," Annual Review of Energy, Volume 4, 1979.

26. Kenneth C. Hoffman and David O. Wood, "Energy System Modelling and Forecasting," Annual Review of Energy, Volume 1, 1976.

27. Jerry Hausman, "Project Independence Report: An Appraisal of US Energy Needs up to 1985," The Bell Journal of Economics and Management, August, 1975, BP.517-51.

28. David Wood and Jerry Hausman, "Energy Demand in the ERDA National R&D Plans, in *An Analysis of the ERDA Plan & Program*, Office of Technology Assessment, U.S. Congress, 1975.

29. John A. Neri, "An Evaluation of Two Alternative Supply Models of Natural Gas," *The Bell Journal of Economics*, Vol. 6, No. 2 (Autumn) 1975.

30. Systems Control, Inc., "Applicability of Brookhaven National Laboratories Energy Models to Electric Utility R & D Planning," Electric Power Research Institute, Palo Alto, California, (EA-807), June, 1978.

31. Charles River Associates, Inc., "Long-range Forecasting Properties of State-of-the-art Models of Demand for Electric Energy," Electric Power Research Institute, Palo Alto, California, (EA-221), December, 1976.

32. Richard L. Gordon, "Economic Analysis of Coal Supply: An Assessment of Existing Studies," Electric Power Research Institute, Palo Alto, California, (EA-496), July, 1977.

33. Milton Holloway, "The Texas National Energy Modelling Project: An Evaluation of EIA's Energy Midrange Forecasting Model," (this proceedings).

34. Saul Gass, "Evaluation of Complex Models," *Comput and Ops Res*, Volume 4, 1977, pp. 27-35.

35. Federal Energy Administration, National Energy Outlook, (FEA-N-75-713) February 1976. Government Printing Office, Washington, D.C., 1976.

36. Peter House and John McLeod, *Large-scale Models for Policy Evaluation*, John Wiley & Sons, 1977.

37. Energy Research & Development Administration, "A National Plan for Energy Research Development and Demonstration: Creating Energy Choices for the Future," (ERDA 76-1) U.S. Government Printing Office, Washington, D.C., 1976.

38. Charles River Associates, Inc., "Review of Large-scale Energy Models," Electric Power Research Institute, Palo Alto, California, (EA-968), January, 1979.

39. Mathematica, Inc., "The Comparative State-of-the-Art Assessment of Gas Supply Modelling," Electric Power Research Institute, Palo Alto, California, (EA-201), February, 1977.

40. Optimal Analysis Company Co., "A Critique of the Federal Energy Administration's Energy Model," report prepared for the House Subcommittee on Energy and Policy, July, 1976.

41. William W. Hogan, "Energy Modelling: Building Understanding for Better Use," presented at The Second Berkeley Symposium on the Systems and Decisions Sciences, Berkeley, California, October, 1978.

42. Saul Gass, "Computer Model Documentation: A Review and an Approach," (NBS special publication 500-39), National Bureau of Standards, Washington, D.C., February, 1979.

43. National Bureau of Standards, "Guidelines for Documentation of Computer Programs and Automated Data Systems," FIPS PUB 38, Washington, D.C., February, 1976.

44. Stanford Institute for Energy Studies, "Stanford-EPRI Workshop for Considering a Forum for the Analysis of Energy Options Through the Use of Models," Electric Power Research Institute, Palo Alto, California (EA-414-SR) May, 1977.

45. James Sweeney, "The Energy Modelling Forum," (this proceedings).

46. Neil L. Goldman, and James Gruhl, "Assessing the ICF Coal and Electric Utilities Model," (this proceedings).

47. James Gruhl and David O. Wood, "Independent Assessment of Complex Models," proceedings of a workshop Validation of Energy Related Mathematical Models, C.R. Deeter and A.A.J. Hoffman (eds.) Forthcoming.

48. Walter Short, John Rambo, and Robert Fuller, "Historical Simulation with the Livermore Economic Modeling System," Laurence Livermore Laboratory, Livermore, CA. Forthcoming.

49. Andrew Ford, Glenn H. Moore, Michael D. McKay, "Sensitivity Analysis of Large Computer Models: A Case Study of the COAL2 National Energy Model," Los Alamos Scientific Laboratory, Los Alamos, New Mexico (LA-7772-MS), April, 1979.

50. Albert J. Eckstein and Dale M. Heien, "A Review of Energy Models with Particular Reference to Employment in Manpower Analysis," Report for the Employment and Training Administration, U.S. Department of Labor, Washington, D.C., March, 1978.

51. Data Resources, Inc., "The Residential Demand for Electricity," Electric Power Research Institute, Palo Alto, California, (EL-235, Volumes 1 and 2), January, 1977.

52. Paul Joskow and Martin Baughman, "The Future of The U.S. Nuclear Energy Industry," Bell Journal, Vol. 7, No. 1, 1976.

53. Professional Audit Review Team, "Activities of the Office of Energy Information and Analysis," Report to the President and the Congress, December 5, 1977.

54. Michael Shaw, John Farquhar, and S. Michael Lutz, "Management and Implementation of PIES Access," and "Recommendations for PIES Access," a report prepared for the Energy Information Administration by Logistics Management Institute, Washington, D.C., March 1978.

55. Michael S. Macrakis (ed), <u>Energy: Demand, Conservation and Institutional Problems</u>, The M.I.T. Press, Cambridge, Massachusetts, 1974.

56. Paul L. Joskow and Martin L. Baughman, "The Future of the U.S. Nuclear Energy Industry," <u>Bell Journal of Economics</u>. Volume 7, number 1, 1976.

57. Edward A. Hudson and Dale W. Jorgensen, "U.S. Energy Policy and Economic Growth, 1975-2000," <u>The Bell Journal of Economics and Management Science</u>, Volume 5, No. 2, (Autumn), 1974.

58. MIT Policy Study Group, <u>Energy Self-Sufficiency: An Economic Evaluation</u>, American Enterprise Institute for Public Policy Research, Washington, D.C., 1974.

59. Nuclear Energy Policy Study Group, <u>Nuclear Power: Issues and Choices</u>, Ballinger Publishing Company, Cambridge, MA, 1977.

60. Energy Policy Project of the Ford Foundation, <u>A Time to Choose: America's Energy Future</u>, Ballinger Publishing Company, Cambridge, MA, 1974.

DISCUSSANT COMMENTS

William W. Hogan
John F. Kennedy School of Government
Harvard University
Cambridge, Massachusetts 02138

I endorse the purposes of this workshop, and I am impressed with the breadth and quality of the speakers that Saul Gass has assembled. As a discussant, I find myself in the happy position of agreeing with the general tenor and substance of the remarks of the speakers who have preceded me. My comments, therefore, are primarily in the form of a few observations and suggestions as to the implications of the chief points of the previous speakers.

The common theme so far is the development of taxonomies of the problems and approaches to the assessment of models. George Lady described the distinctions among the concepts of model verification, validation, and ventilation, emphasizing the critical importance of understanding the objectives of any examination of a model and its uses. Dave Wood expanded this taxonomy to distinguish between the types of models: policy research versus policy analysis models. It is clear that models and model uses are heterogeneous, and this diversity must be recognized in the design and use of the model assessments. At one end of the spectrum, we have models constructed to examine the state of nature and to test hypotheses, and the parallel model assessments are designed for the benefit of scientists and model builders. At the other extreme we have models constructed to deal with the messy policy problems immediately at hand, and the evaluation structure and standards must be adapted to the model user and the decision maker.

This mapping of the terrain is a primitive but crucial step for the development of a model assessment process. And it points to a number of immediate implications:

1. The process is expensive. Assessing a model and its uses can be an activity on the same scale or larger than the effort of constructing the model in the first place. If this essential component of an analytical process is to develop and mature, then it must be funded on a scale commensurate with its benefits and its difficulty. The sponsors to date have been generous in their support of the early phases of model assessment processes, but it is clear that the actual costs of model assessments have greatly exceeded our earlier expectations. It is essential that future model assessment efforts be put on a stable and professional basis.

2. We should not rush too far ahead, to the setting of standards, too fast. It is clear that we do not know much about the model assessment process, and a great deal more experimentation and research will be necessary before we can approach the panacea of validated models that can be certified for third-party use. The National Bureau of Standards is to be complimented for organizing this workshop, but they should not be misled to thinking they are near the establishment of modeling standards.

3. The terrain we understand the least is where the models and analytical information are used by decision makers and how to help them and the modelers in developing credible modeling tools. At the scientific end, the research models have analogies from the hard sciences, with standards of validation and replication that can be adapted to the model assessment process. For policy analysis, however, they are not quite sure what the game is, but the analogy may be closer to the work of lawyers in search for truth through the advocacy process. This behavioral research into the attitudes of model users seems essential. For example, the work of James McKenney of Harvard reports results which can classify decision makers as either analytical or intuitive. The types of models and the types of validation approaches that would be appropriate for the intuitive decision maker are quite different from those necessary for the analytical person. Yet the analytical approach is the typical setting for the modeler. More research, integrating the behavior of the users with the technical characteristics of the models, seems necessary if we are to make progress in understanding how models will be used. We need innovative work here.

I am pleased that we are making such rapid progress in defining and investigating the model assessment process. I hope the remainder of the workshop will provide further examples of the process, to help improve the assessments without having the assessment detract from the application of the model or the benefits that can accrue from the construction of new and better systems.

James L. Sweeney

Energy Modeling Forum*
Terman Engineering Center
Stanford University
Stanford, California 94305

INTRODUCTION

In recent years, especially since the oil embargo of 1973, there has been widespread development of energy models in the executive and legislative branches of government, universities, industry, research institutes, consulting companies, and commercial establishments. Unfortunately, the ability to utilize the models effectively for energy policymaking and planning has not kept pace with this development. The gap between modelers and potential users of models is large and pervasive. Heightened concern about energy problems coupled with the proliferation of analytical tools for addressing these problems has created both the need and the opportunity for bridging the gap. Finding ways to improve communication between model developers and model users has become an active area of investigation and innovation [5].

The Energy Modeling Forum (EMF) has been one response to this situation. The EMF seeks to improve the use and usefulness of energy models in the study of important energy issues. Sponsored by the Electric Power Research Institute (EPRI), the EMF is headquartered at Stanford University within the Institute for Energy Studies and the Departments of Engineering-Economic Systems and Operations Research. The Forum operates through a series of ad hoc working groups consisting of roughly equal numbers of energy modelers and potential energy model users. Each working group focuses on an issue or set of closely related issues important to energy policymaking or planning and to which existing energy models can be applied. The group designs, implements, interprets, and communicates a set of tests designed to illuminate the basic structure and behavior of the models. The issues addressed by the group thus provide a forum to compare and contrast the various models, identifying their capabilities and limitations. At the same time, the issue focus assures that the policy-relevant implications of the various models are developed and communicated.

The basic goal of the Energy Modeling Forum — to improve the use and usefulness of energy models in the study of important energy issues — entails a number of subgoals, some competing, some complementary.

The first set of goals relates to comparison of models in order to improve understanding of their limitations and capabilities:

- To identify and compare critical elements of existing energy models and to illuminate their major strengths and weaknesses.

- To cast light on key modeling issues so as to afford a greater understanding of alternative modeling approaches; and

- To provide guidance for the improvement, linkage, and extension of energy models and to establish priorities for new modeling research.

Within the first subgoal, the word "existing" should be emphasized. The Forum does not attempt to create new models but rather to identify and compare existing models and modeling approaches. The second subgoal entails a generalization from the specifics of individual models to issues having general relevance to a broad class of existing or potential models.

The third subgoal is an overstatement of the activities of the Forum to date. The EMF does not establish priorities for new modeling research. However, the results of the studies do suggest priority research areas for model improvement.

These subgoals focus on models, the modeling process, and the supporting research. The second set of subgoals, on the other hand, focuses on improving information for energy policy and planning through the appropriate use of models. These subgoals are more concerned with the information available through the models than with the models themselves:

- To use major energy models to sharpen insights, improve understanding, and explore the implications of selected energy decisions and scenarios; and

- To broadly disseminate information about possible energy futures and the impacts of various energy actions on those futures.

These goals to a large extent are more oriented towards people and their use of models than towards the models themselves. This point is made quite forcefully by Martin Greenberger, in a paper prepared for this volume. The goal is to sharpen insights and understanding about the impacts of various energy decisions, but perhaps more importantly to improve understanding about the use of models as tools for

sharpening insights and improving understanding. This goal
of improvement in understanding is held both for model users
and modelers, groups which often do not adequately
communicate with one another.

A set of design principles guides Energy Modeling Forum
activities in pursuit of these goals [9]:

- User Orientation. The EMF should work to improve
  the use and usefulness of energy models, approaching
  the studies from the user perspective and
  maintaining an active user involvement.

- Model Comparison. The EMF studies should compare
  the capabilities and limitations of many models,
  and these comparisons should be descriptive rather
  than normative. This is a unique contribution that
  the EMF can make, and it avoids the difficult
  problem of model validation.

- Issue Focus. For the general model user, abstract
  model comparison should be conducted in the context
  of the application to an important energy issue.
  This will provide a direction and discipline for
  the model tests.

- Broad Participation. The communication objectives
  of the EMF are best served if there is a wide
  participation in the selection of study topics, the
  formation of the working groups, and the
  dissemination of the study results.

- Decentralized Analysis. Existing energy models are
  often complex and require skillful application by
  the model developer. Despite the inherent
  advantages of third-party analysis, the EMF must
  rely on model tests as reported by the individual
  research group [9].

In general, the studies to date have conformed well to the
design principles and guidelines. All studies, with the
possible exception of EMF 4, have maintained an active user
involvement, including users from government agencies,
private sector corporations, research institutes, and
universities. Each of the studies has included model
comparisons, using between 6 and 18 models, with 10 being
the median number addressed in any one study. Each had a
strong issue focus although the immediacy of the issues
varied significantly among the studies. The two final
design principles, broad participation and decentralized
analysis, have been fully met in each of the studies.

The current organizational structure of the EMF is illustrated in Figure 1. The Senior Advisory Panel, the working groups, the EPRI staff, and the EMF staff interact with one another and with the broad community of energy modelers and potential model users, as well.

The heart of the EMF consists of the ad hoc working groups of about 35 members each. The working group chairman and the issue to be studied are selected before the formation of the working group. Each working group, composed of volunteer participants, with a balanced representation of model users and model developers, is organized around a specific energy issue to ensure both the proper representation of relevant models and participant interest in the policy or planning issues addressed. The chairman selects members with the goal of obtaining a working group which is diversified geographically, institutionally, and philosophically. Observers and other closely follow the working groups' progress; the working groups, in turn, rotate as new issues are selected. To date, over 250 people have been involved in the five studies to date, and more than 100 might be active in any year.

At each stage in this process, the EMF is assisted by the Senior Advisory Panel. This group, chaired by Charles Hitch of Resources for the Future, Inc., is composed of senior energy decision makers (see Appendix for a list of the membership) who represent the ideal target audience for the EMF studies. The Panel helps maintain the necessary broad participation and user orientation to assure the value and immediate relevance of the working group topics. The Panel meets annually and provides necessary advice throughout the year. Its functions are primarily fourfold:

- o to suggest appropriate study topics and to critique prominent study proposals so as to provide a sense of priority,

- o to suggest and possibly assist in recruiting appropriate working group chairmen and members,

- o to critique the working group's final report in draft form both for substance and presentation, and

- o to help disseminate the results of the studies.

The overall planning, coordination of daily operations, and administration of the Energy Modeling Forum are handled by the EMF staff, supervised by an Executive Director (William Hogan from September 1976 through August 1978 followed by James Sweeney). Located at Stanford University, the EMF staff is affiliated with the Stanford Institute for Energy Studies and the Departments of Engineering-Economic Systems and Operations Research. The staff (see Appendix for a listing of staff members) provides support for the Senior Advisory Panel in the development and selection of issues for future topics, recruits the working group chairman, assists the working group chairman in organizing a study,
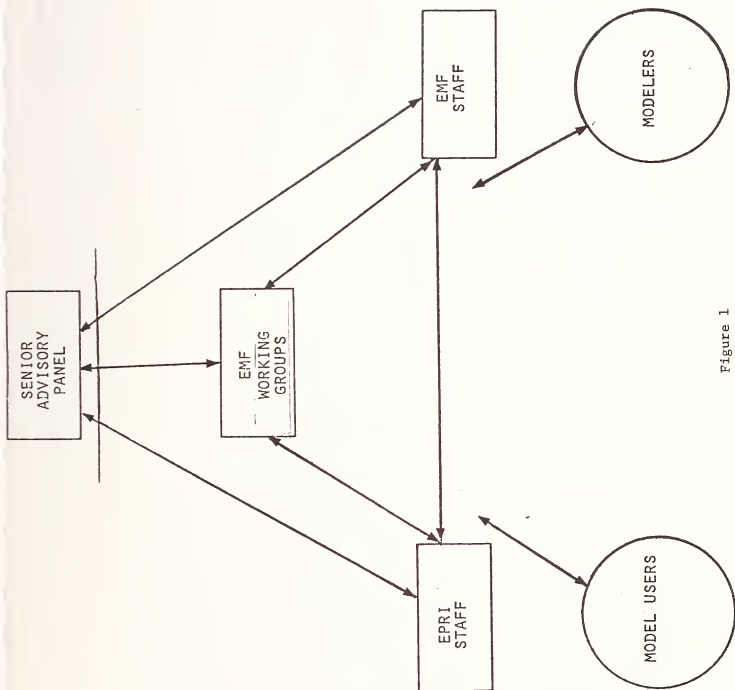
Figure 1

participates both as members of the working group and staff to the group, and publishes the final working group reports.

The communication function of the EMF is enhanced by close ties maintained among the various participants in the Forum. The Senior Advisory Panel includes EPRI representation. Meetings of the Senior Advisory Panel are normally attended by the Executive Director of the EMF, at least one working group Chairman, and by several EPRI staff members. Energy Modeling Forum working groups include EMF staff members as well as EPRI staff members particularly knowledgeable in the area being addressed. Close coordination between the EMF staff and the EPRI staff is maintained throughout all phases of the Forum. Some of the working group members or chairmen were initially proposed by Senior Advisory Panel members. The community of energy modelers and potential energy model users interacts broadly with the entire process by participating directly in working groups as members or observers, by membership on the Senior Advisory Panel, or by a one- to three-year position with the EMF staff. This community also suggests appropriate study topics, models to be considered, and issues to be addressed, and maintains informal communication with EMF groups.

THE EMF PROCESS

A typical study cycle begins with a broad call to modelers
and potential model users to assist in identifying potential
study areas, moves through phases of working group
organization, intense modeling activities, result
interpretation, and the writing and publication of the
report. The complete process may take as long as a year and
a half, involving typically three to four working group
meetings, spaced about three months apart. Publication is
normally followed by an indefinite period of publicizing the
study and development of applications for the work. Table 1
illustrates the various phases of studies and indicates
which groups typically are actively involved in various
phases of a study.

The process of selecting a study topic involves a wide range
of participants. Initial identification of potential topics
is accomplished by collecting ideas offered by many people.
These suggestions are distilled to a dozen or so major
potential study areas to be considered by the Senior
Advisory Panel. The Panel in discussing the issues provides
a sense of priority from a user perspective as well as
providing suggestions for specific issues within the general
areas. Additional preliminary exploration and issue
identification of high priority topic areas follow. This
activity is coordinated by the EMF staff but involves
participation of the community by energy modelers and model
users.

This process results in the selection of a topic and a
chairman to direct the working group. Concurrently, the
chairman, in coordination with the EMF staff, selects the
specific project to be undertaken. This simultaneity
assures that the chairman is not only directing a study on a
topic within his area of expertise and interest but also one
which seems feasible in light of the existing energy models
and the limitations on the current state of the art.

A working group is recruited primarily by the chairman and
the EMF staff, with the working group chairman guiding the
selection process. This phase, along with the first step of
selecting the working group chairman, is critical to the
study's success. The value of the process and the final
report is dependent upon a strong, knowledgeable, diverse
working group whose members are familiar with nuances of
policy issues and policy models and devoted to improving
applications of models to policy and planning issues.

Once the working group is organized and holds its first
meeting, complete responsibility for the conduct of the
study is vested in the group.

The working group selects the models to be run. Normally,
there is agreement that each existing current model
represented by a working group member can participate.
However, the working group may identify and recruit
additional modelers.

During the first meeting, group members identify the most

Table 1

PARTICIPATION IN PHASES OF EMF STUDIES

Study Phases	Senior Advisory Panel	EPRI	Working Group Chairman	Working Group Group	EMF Staff	Others
Identification of Potential Area	X	X			X	X
Consideration by Senior Advisory Panel	X	X			X	X
Preliminary Exploration and Issue Definition	X	X			X	X
General Topic Area Selection	X				X	
Working Group Chairman Recruiting	X	X			X	X
Specific Topic Selection			X		X	
Working Group Recruiting	X	X	X		X	X
Model Selection			X	X	X	
Issues Identification			X	X		
Scenario Specification			X	X	X	
Selection of Output Variables			X	X	X	
Running of Models				X		
Displays of Results				X	X	
Model Comparisons			X	X	X	
Critique and Interpretation of Runs			X	X	X	
Write Executive Summary			X	X		
Write Summary Report			X	X	X	
Write Appendixes			X	X	X	
Critique Report	X		X	X		
Publication of Report		X			X	
Publicizing Study and Applications	X	X	X	X	X	X

important energy policy or planning issues to be addressed
and those that cannot be addressed. The participation of
modelers and model users is critical at this point.
Capabilities and limitations of the various models, in
addition to priorities for further model development, soon
become apparent. In general, the goal here is to focus upon
the most important issues and to test the capabilities of
the various models to address these issues. This is to be
contrasted with what can easily happen in practice: the
issues to be dealt with are determined by the particular
strengths of the individual models, with secondary
consideration given to the importance of the issue for
policy and planning purposes.

Study issues normally relate to informational questions
significant to energy policy and planning. For example, the
second study asks to what extent alternative environmental
restrictions influence both the rate of the transition back
to coal and the regional distribution of the growth. The
fifth study asks how the U.S. supply of oil or natural gas
will be influenced by domestic price controls.

Alternatively, the study question may be a modeling or
forecasting issue. The third study asks how significant
changes in the price of electricity will influence
consumption. Differences in the answers among models seem
to depend upon the geographic scope of the data base used
with the models. Thus, the issue becomes: to what extent
could combined historical data from many regions be used to
improve the estimation of parameters for a model applicable
to a single utility service area?

Once the issues are identified, scenarios are generated to
capture the essential features of the issue being
considered. The scenario specification includes a set of
standardized input assumptions, some of which are changed
systematically among the scenarios to test the models and to
provide information potentially useful for policy and
planning purposes. In general, each modeler is asked to run
each of the scenarios with the standardized assumptions.

Differences in the models become apparent. The input data
for one model may be the output of another. Some input
variables may not be included in a model, even though they
are believed by working group members to be particularly
significant in influencing the projections. Some scenarios
simply cannot be addressed with a particular model except in
an ad hoc fashion, requiring extensive off-line manipulation
by the analyst. Documentation of these differences is
important when comparing the results of the various models
and the models themselves.

The definition of scenarios is accompanied by a selection of
output variables to be reported by each modeler. These are
normally selected by the working group in order to provide
projections of those variables that are most meaningful for
policy and planning purposes and to give the working group
members an opportunity to observe the inner workings of a
model at several critical points. For example, in the fifth
study, output variables to be reported will include measures

of drilling activity, quantities of oil and gas discovered, reserves of oil and gas, and production rates for a selected set of years. Whether or not individual models can calculate these intermediate variables is often important in communicating their capabilities, limitations, and dependability. Comparisons of model behavior at several observation points provide the raw material for group discussions and report writing.

Each model is run by its key developer or a close colleague. Third-party model operation is not routine. There are advantages in third-party model operation, but these are outweighed by the practical advantages of the modelers running their own systems. The developer may best understand the limits of applicability of the model. Moreover, the modeler knows which sets of equations or data within the model must be modified to examine specific scenarios. Having worked extensively with a given system, (s)he can make runs without undertaking the enormous learning cost required for third-party analysis. In addition, decentralized model operation supports the goal of enhanced communication among the model users and model developers.

Displays of model outputs are designed so as to facilitate interpretation and comparison. Graphic comparisons are prepared by the EMF staff for use by the working group. The displays of the outputs, if done creatively, can help in interpreting the comparative behavior of the individual models. If done poorly, the display of results can hide more information than it communicates. Thus, this task is far more critical than may be apparent.

A large portion of working group time is devoted to the critique and interpretation of the runs. Differences among models provide an opportunity and a motivation for explaining why the different results occur. Discrepancies in results may point to fundamental differences in model structure, model parameters, basic data utilized, or perceptions about the direct implications of scenario assumptions. Divergence in model results normally leads to creative tensions among the modelers with each trying to understand why his model differs from the others. One motivation is to improve the model, if appropriate. Another is to show why one model's answers may be more dependable than those of another model. Divergences in results is a strong motivating force which leads to important understandings about the fundamental model differences relevant for policy or planning purposes, the areas of uncertainty in knowledge about the world, and the significant areas of research potential. The process has resulted in the revision of a model during the study to account for implementation problems not initially perceived by the model's developer.

These working group discussions enhance the model users' insights about the policy issues and suggest distinctions among policy options that may not be apparent on the surface. For example, some policy options may increase the cost of producing a given amount of energy, e.g.,

restrictions on types of coal that can be used for electricity generation or restrictions against the use of nuclear energy. Others, however, may simply influence the price of energy without influencing its cost, e.g., an energy tax that is redistributed through the U.S. Treasury. The first class will have a far greater impact on economic growth than will the second class of options [10]. What may seem to be a subtle distinction in policy options may profoundly impact effects of the various options on economic growth.

The working group report is characteristically written and published in two separate components: a relatively short summary, approximately 30 pages long, directed at a broad audience and a much longer series of supporting documents or chapters aimed at a smaller, more technical audience.

The summary explains the major commonalities and differences in the models, provides answers (to the extent possible) to the issues raised, identifies limitations of the analysis, and presents recommendations developed by the group. In writing this report communication is emphasized; the report is intended to be jargon-free and accessible to nontechnical readers. A two- to three-page executive summary encapsulates the key conclusions of the report.

The series of supporting documents varies significantly from study to study. Generally included, however, are descriptions of the individual models, comparisons of the models, a simplified framework for both comparing the models and communicating an intuitive understanding of the results, a detailed description of the scenarios, detailed results from each model with comparative graphics, and a set of technical papers discussing more deeply any modeling and analysis issues that may have surfaced during the study.

The working group report provides one communication vehicle for disseminating the results of individual studies. Other mechanisms also are used. The Senior Advisory Panel is briefed on the report and members have played significant roles in communicating the results. Working group observers bring insights back to their respective organizations and help to disseminate the study's findings more broadly. Working group members typically make seminar or conference presentations based upon the study. EPRI, as well as the EMF, publishes the report and facilitates its distribution. Thus, many individuals help publicize the study. The study belongs fundamentally to the working group and the communication of results relies heavily on study group members actively publicizing the results.

Although the EMF reports, individual participant presentations, and other vehicles are used to communicate the results of the study, much of its benefit is not easily transferred to nonparticipants. A major EMF focus is on how people can use models more effectively, but effective use is a skill, learned like any other skill. While an EMF study can help modelers and model users in the difficult, artistic process of utilizing models for addressing real, complex issues, the skills and insights gained often cannot be fully

transferred beyond the participants, except at extremely high cost. Hence, formal communication of results beyond the working group is an important product of EMF studies, but clearly not the only product.

EMF STUDIES

The products of the Forum* - comparative studies of
significant energy issues - include two completed studies,
"Energy and the Economy" and "Coal in Transition:
1980-2000," and one virtually completed study, "Electric
Load Forecasting: Probing the Issues with Models."
"Aggregate Energy Demand Elasticities" is well under way;
and a fifth study, "U.S. Oil and Gas Supply," is just
beginning.

Table 2 summarizes the progress of the various studies as of
March 1, 1979.

Two other study topics identified by the Senior Advisory
Panel as high priority areas, "Energy and the Environment"
and "World Oil Supply, Demand and Prices," are listed in
Table 2. Although at this time it is uncertain whether
either will be chosen for future study, the EMF staff is
currently in the process of preliminary exploration and
issue definition.

This section describes the studies, focusing primary
attention on the study process and on the contribution of
each to the evolution of the forum.

EMF 1: ENERGY AND THE ECONOMY

The Forum project was initiated in 1976 with a study
designed to demonstrate the research concept. The working
group used six models to study the nature and strength of
the feedback from the energy sector to the aggregate
economy, isolating the key factors determining the effect of
energy system changes on the long-run economic growth. The
results demonstrate the importance of the value share of
energy in the economy, the flexibility in substituting other
inputs for energy use, and the link between productivity and
capital formation in explaining the behavior of the models
[4].

A group of 30 model users and developers conducted the
study. Because of the experimental nature of the project,
William Hogan served as working group chairman as well as
EMF Executive Director.

The six models each explicitly represented the
energy-economic linkage. Each model was for the full U.S.
economy, and each was judged appropriate for long-run issues
but not for short-run issues. Common scenarios were
constructed by standardizing many input variables. The
working group then sought to explain the common results or
the causes of model differences. This comparison process
was facilitated by the high degree of commonality among the
various models.

The key comparative results of the study were estimates of
the aggregate elasticity of substitution** implicit in the
participating models. This parameter was shown in the

Table 2

PHASES OF EMF STUDIES[a]

Phases	EMF 1	EMF 2	EMF 3	EMF 4	EMF 5	Energy/ Envir.?	World Oil?
Identification of Potential Area	C	C	C	C	C	C	C
Consideration by Senior Advisory Panel	C	C	C	C	C	C	C
Preliminary Exploration and Issue Definition	C	C	C	C	C	●	●
General Topic Area Selection	C	C	C	C	C		
Working Group Chairman Recruiting	C	C	C	C	C		
Specific Topic Selection	C	C	C	C	C		
Working Group Recruiting	C	C	C	C	C		
Model Selection	C	C	C	C	●		
Issues Identification	C	C	C	C	●		
Scenario Specification	C	C	C	C	●		
Selection of Output Variables	C	C	C	C	●		
Running of Models	C	C	C	●			
Graphical Displays of Results	C	C	C	●			
Model Comparisons	C	C	C	●	●		
Critique and Interpretation of Runs	C	C	C	●			
Write Executive Summary	C	C	C	●			
Write Summary Report	C	C	●	●	●		
Write Appendixes	C	C	●	●			
Critique Report	C	C	●				
Publication of Report	C	C					
Publicizing Study and Applications	●	●					

C – Completed as of February 28, 1979
● – Ongoing as of February 28, 1979

[a] EMF studies: EMF 1: Energy and the Economy
EMF 2: Coal in Transition: 1980–2000
EMF 3: Electric Load Forecasting: Probing the Issues with Models
EMF 4: Aggregate Demand Elasticities

report to be one of the key determinants of the strength of the link between energy and the economy.

The first EMF study contributed importantly to the current structure of the EMF process. In particular, the study involved considerable participation of the model-using community, which resulted in careful attention to specifying limitations of the models in studying the energy-economy issue. Additionally, the study had a strong issue focus. Although some questioned the direct applicability of the study's results to the evaluation of the energy policy options available to the federal government, it served to educate many policymakers about the magnitude of the relevant trade-offs.

Despite its positive contributions, the first EMF study suffered from several problems that had plagued previous model comparison studies. The group was often torn between the sometimes conflicting goals of policy analysis and model comparison. The study suffered from a lack of visibility; and it had a distinctly academic flavor.

The issue focus of the study was intended to aid in the model comparison. However, this focus implied to some working group members that policy recommendations could be drawn from the study's conclusions. Other study participants contended that they would have conducted a policy study differently from a model comparison study and, therefore, argued against the development of policy recommendations.

Despite the active participation of the user community in the study, it suffered from a lack of visibility at the highest levels of government and industry. The perceived lack of visibility of the first study led to decision to institute the Senior Advisory Panel, discussed previously.

EMF 2: COAL IN TRANSITION: 1980-2000

A second EMF working group, organized in July 1977, compared 10 different models in the analysis of coal production, distribution, and utilization. The report documents the greater importance of coal demand issues relative to supply issues and describes various insights into the level and composition of future coal output gleaned from the models' results. Emphasis is placed upon the sensitivity of patterns of future coal use to changes in regional economic conditions and standards on allowable emissions [2].

This study differed in several respects from the first. The Executive Director did not serve as working group chairman. Rather, Dr. David Sternlight, Chief Economist at the Atlantic-Richfield Company (ARCO), served as chairman. His industrial affiliation, user perspective, and previous energy policy analysis and modeling experiences proved to be invaluable. This choice allowed the complementary talents of the working group chairman and the EMF Executive Director to be jointly applied in the leadership of the study.

The models in the second study differed in scope from those

in the first study. Three types of models were employed:
energy sector models with significant coal detail; models of
coal supply, transport, and demand; and a resource planning
model. It was now more difficult to standardize
assumptions: e.g., the energy sector models used exogenous
projections of aggregate energy demand as inputs, whereas
the coal and facilities planning models required exogenous
demand projections for electricity consumption and for
nonutility coal demand. Model comparisons were conducted
not only in the parallel mode, as in the previous study, but
also in a complementary mode, with different information
developed by different models. In the complementary mode,
the results of one model, for example, a detailed energy
sector model, could be scrutinized by use of another, more
disaggregated model - for example, a resource planning
model.

The conflict between the model comparison and policy
analysis goals that surfaced in the first study was again
apparent in "Coal in Transition." Once more, one group
resisted the notion that the study would make policy
recommendations while another urged the development of such
recommendations.

EMF 3: ELECTRIC LOAD FORECASTING: PROBING THE ISSUES WITH
MODELS

The third study group was designed to help electric
utilities deal with the new complexities and uncertainties
of electric load forecasting. This group examined the use
of 10 current models in forecasting electric loads. The
experiments identified and illuminated prominent load
forecasting issues and improved understanding of the models'
capabilities and limitations [3].

Bernard Cherry, Vice President for Corporate Planning of the
General Public Utilities Service Corporation, served as
working group chairman, thus repeating the second study's
successful practice of having a working group chairman from
the relevant industry. Again, the chairman's problem
orientation provided critical guidance and discipline for
the study.

The study contrasted with the previous two in that it
involved models with differing geographical coverages. Most
of the models were used for a particular utility's load
forecasting and considered only that utility's region. This
was a critical issue in the study because regional
differences made it undesirable, if not impossible, to
standardize the inputs to the models. The standardization
problem was circumvented by allowing each modeler to specify
a "best information" base case. Input parameter variations
between scenarios were specified in percentage terms, and
scenario results were examined in terms of percentage
differences in values of output variables.

The variance in scope of the models provided observations of
differences in model behavior. In particular, for given
percentage changes in electricity prices and in competitive

fuel prices, applications of the few nationally oriented models in this study showed a larger proportional impact on electricity consumption than did the utility region models. As the differences were addressed, the striking dissimilarities led to debate among the modelers. The observation led one group to conclude that combined data from many utility service regions could be used to more accurately estimate parameters in the demand models than would be possible by using data from only a single utility region. Another group felt that combining this data would reduce the quality of the estimates because of the great differences among regions. Although consensus was never achieved, the interchanges led some participants to consider the benefits of changing utility forecasting practice by estimating parameters on combined data, and it led other participants to further critical examination of the econometric foundations for pooling data from many regions.

Many participants in the third EMF working group initially questioned the potential value of the study and worried about the possibility of misuse of the results. By the time the study was completed, however, the participants were calling for continuation and expansion of this type of activity. Strong support by the working group members helped the Electric Power Research Institute launch its Utility Modeling Forum (UMF), a comparative analysis project focusing specifically on the problems of the electric power industry, in the autumn of 1978.

EMF 4: AGGREGATE ENERGY DEMAND ELASTICITIES

The fourth working group is conducting a specialized test of the aggregate price elasticity of demand implicit in the participating energy models. Eighteen models were run under nine scenarios testing the models' responses to variations in the prices of oil and gas, coal, and other energy sources. Interpretations of model runs and conclusions are still under heated debate [1].

This study is somewhat different from the earlier studies. Motivated partly by the EMF 1 conclusion that the aggregate elasticity of substitution is a critical determinant of the link between energy and the economy, the Senior Advisory Panel recommended that the EMF perform further experiments to improve the precision and level of confidence in the elasticity estimates. They felt, however, that the experiment would be too technical and not tied closely enough to specific policy issues to warrant formation of an EMF working group. Thus, during early 1978, the EMF staff, with the aid of many outside experts, designed an experiment to estimate the aggregate demand elasticity implicit in energy models.

By late 1978, 18 models had executed the experiment. Interest in the study's results had escalated to the point where a face-to-face meeting of the study participants and interested observers was deemed desirable. Therefore, a working group of approximately 40 people, predominantly model builders, was formed. Hogan, by then at Harvard University, became working group chairman, thus providing

continuity in the transition to James Sweeney's tenure as Executive Director. One meeting of the working group has been held to date, and a second meeting probably will be necessary.

The scope of the models of the fourth working group varied greatly. There were self-standing U.S. models and U.S. models as components of international energy demand models. Some were highly aggregated; others were disaggregated by fuels. Some models included the energy sector embedded within the entire economy. Some examined the aggregate of all energy consuming sectors, while others were disaggregated by sector or represented only a single energy-consuming sector.

Ongoing work is proceeding on several fronts: an examination of the properties of the aggregate elasticities, development of a taxonomy of models, and a characterization of the uncertainty associated with the forecasts.

The experiment has measured the aggregate elasticity of energy demand implicit in the participating models, but properties of that aggregate measure are not yet fully understood. Do the measured elasticities depend significantly upon what indices are used for the aggregation? How sensitive is the aggregate elasticity to various combinations of changes in the energy prices? If some energy prices decline while others increase, is the aggregate elasticity the same as that obtained when all prices increase proportionately? To what extent does the specific trajectory of price changes influence the measured elasticity in the models?

A taxonomy of models is being addressed by working group members in order to reconcile, to the greatest extent possible, the elasticities obtained using the different methodologies. Issues of uncertainty are being adddressed: a method is being devised for characterizing models which is not simply in terms of implicit elasticity but also in terms of the variance associated with that elasticity.

Although it has not yet been completed, the fourth EMF study has raised the question of whether a technical comparison study can best be conducted apart from the model-using community. However, the technical comparison must be focused on one or two relationships widely believed to be significant. The ultimate utility of the study will depend upon our ability to widely communicate the results effectively to the model-using community as well as to modelers.

EMF 5: U.S. OIL AND GAS SUPPLY

The fifth working group held its first meeting in January 1979. The group plans to examine the effects on domestically produced oil and gas of alternative world prices for oil, domestic prices for natural gas, oil price controls, alternative federal leasing rates, price controls, surprises in price trajectories, changes in the tax structure, and alternative assumptions about the geological

resource base.

Consistent with previous studies, Ben Ball, the chairman (currently an Adjunct Professor of Management and Engineering at the MIT Energy Laboratory and formerly a Vice President of Gulf Oil Corporation), has fundamentally a user perspective.

It is too early to predict the progress of the study. One difference, however, from past studies is in the attention placed on model examinations and comparisons. It is hoped that deeper model assessments and comparisons will be possible. Several steps are being undertaken to accomplish this goal. Energy Modeling Forum staff members have drafted and made available to working group members a comparison of major oil and gas supply models. This model comparison paper is delving more deeply into the internal structures of the different models than have previous such EMF papers.

The working group has decided to examine the behavior of the models at several critical points within their structure. For each scenario, all modelers will report drilling activity, reserve additions, reserves, production, and cumulative production over time, for oil, associated gas, and nonassociated gas (for several geographical regions). It is believed that examination of the behavior of these variables (and variables such as the reserve-to-production ratio) and in reponse to the assumed changes will lead to many insights into the capabilities and limitations of the various models.

Several recently initiated projects are closely related to the Energy Modeling Forum. The previously mentioned Utility Modeling Forum (UMF) plans to conduct an ambitious series of studies focused entirely within the utility modeling area. Part of the motivation for the UMF was the EMF 3 success and the positive response of EMF 3 working group members. This project, sponsored by EPRI, and administered by Booz, Allen & Hamilton, Incorporated, has recently held its first meeting.

An activity complementary to the Energy Modeling Forum is the Energy Policy Analysis Forum under the direction of Kenneth Hoffmann of Brookhaven National Laboratory. This forum involves high-level Department of Energy analysts and outside modelers and analysts. Conducted as a continuing seminar, its goal is to identify analytical capabilities dealing with important energy policy issues being considered by the Department of Energy. Individuals in this seminar are expected to discuss a much larger number of issues than can be addressed within the EMF format. The seminar, however, will be limited to the identification of analytical capabilities, and will not conduct its studies in the depth possible in the EMF. Close coordination between the EMF and the Energy Policy Analysis Forum is being maintained.

A third related activity is being contemplated by the Solar Energy Research Institute. This organization may launch an EMF-like comparative model study of alternative new technology diffusion models. This effort would help to identify methodologies that could be employed to model the rates of introduction of solar energy technologies.

A "Model Verification and Assessment" project (discussed in greater depth by other authors contributing to this volume) was established by EPRI at Massachusetts Institute of Technology in 1977. It complements the EMF by going more deeply into the testing and appraisal of individual models. The assessment project has the dual purposes of: (1) developing procedures and methodologies for in-depth assessment, and (2) applying these assessment procedures to individual energy models [8].

While each of these various activities performs somewhat different functions, lessons learned in any one can be valuable to the others. It is hoped and anticipated that extensive sharing of information naturally will occur.

ISSUES FOR THE FUTURE

Unresolved issues to be addressed include:

- o the appropriate trade-off between model comparisons and policy analyses,

- o the extent to which the EMF should conduct model assessments and evaluations,

- o the appropriate role of the EMF in an academic institution such as Stanford University, and

- o the extent to which study participants should be compensated for their time and computer expenses.

A tension keenly felt in the first two studies and anticipated in the fifth is the appropriate trade-off between model comparisons and policy analyses. A typical working group includes a mix of people, some primarily concerned with modeling and some concerned primarily with using information for policy and planning purposes. Indeed a major goal of the project is improved communication between these two groups. While these two activities are complementary in many respects, in other respects they conflict. The model comparisons considerably improve the policy analyses by refining the quality and the reliability of the information developed by using models and by indicating key areas of uncertainty. The policy analyses enhance the relevance of the model comparisons by structuring the comparisons to focus attention on similarities and differences most relevant for policy planning issues.

The first area of conflict between these two goals comes in the choice of scenarios. Because of limited time and resources, only a small number of scenarios can be structured, implemented, and interpreted by the group. Since some scenarios are most useful for model comparisons while others would be most useful for policy analyses, the selection of scenarios represents implicitly or explicitly a choice between model comparisons and policy analyses. The model users would have little interest in a study devoid of policy implications. However, the model developers are anxious to examine the comparative advantages of their systems. While efforts have been made to choose scenarios to satisfy both purposes, the tension has been strongly communicated.

The tension between goals also is felt in subsequent interpretations of the results of the model runs, with conflicts between the allocation of group time for model comparisons and policy analyses. It has been generally maintained that users would see little value in detailed, jargon-ridden debates on the equations and data embedded in the models. On the other hand, without such debates, in-depth model comparisons may be impossible. One proposed solution is to extend one or more working group meetings by a day to allow the modelers, and others desiring to

participate, an opportunity for in-depth model-oriented discussions. Another possible solution would be for the EMF staff to draft more complete model comparison papers and to allow some of the debate to proceed by mail and telephone, directed toward improving the draft.

In developing the final report, the tension over goals is also apparent. The various goals imply different themes and formats for the report. The solution to date has been to write a summary report which focuses on the policy analysis and on the capabilities and limitations of the models as a class. Detailed comparisons of the individual models and the various modeling approaches then appear in a longer report. While this compromise gives weight to both goals, it tends to downplay the model comparison objective.

The issue of policy analysis versus model comparison may never be resolved, but this may be a sign of health. The tension between the goals provides opportunities for working group members to concentrate their own efforts primarily on aspects of the study most relevant to them and thereby helps to improve the overall study quality.

A closely related issue concerns the depth to which EMF should conduct model assessments and evaluations as opposed to simply model comparisons. The EMF recently has been criticized, particularly by the academic community, for its lack of critical review of the participating models.

The Forum has focused attention on the "ventilation" of models, the simple examination and explanation of their behavior [7]. Of course, this step logically must precede evaluation or assessment. This comparative study of the behavior of a number of models allows consequent work to identify differences stemming from data differences, structural differences, differences in explicit assumptions, and, often most importantly, differences in implicit assumptions or world view held by the developers. Although evaluation per se has not been conducted by the Forum, the differences identified through the comparisons provide an improved basis for individuals to make their own evaluations of the models.

There are several reasons why the Forum up to this point has not conducted in-depth evaluations. First is the question of the extent to which objective comparisons are possible. Of course, some aspects of assessment can be conducted objectively. One could examine whether the computer code was written as the developer intended or could attempt to replicate the underlying econometrics. Activities of this sort are in fact being conducted by the MIT Model Assessment Laboratory [8]. Some assessments, however, cannot yet be objective with the current state of the art but are based upon subjective peer review judgments. Econometric evidence, for example, can be viewed differently by various professionals. Even more difficult is sorting through and assessing the implicit assumptions and the world view incorporated in the model. Evaluating which implicit assumptions and world view are more nearly correct cannot be done objectively. Individuals, however, can make their own

judgments on a subjective basis if the behavior of the models is clearly communicated. This individual, subjective process is facilitated by the Forum studies.

Furthermore, early experience at the Model Assessment Laboratory has demonstrated that a credible in-depth, hands-on, third-party review of a single model can require resources comparable to a full EMF study. A comparative in-depth assessment of, perhaps, 10 models could require an order of magnitude more resources. Even if desirable, such a process is impossible for the Energy Modeling Forum as it is faced with a limited budget.

Another difficulty with model evaluation is that different models may be particularly useful for different purposes. The type of model useful for forecasting the consumption of gasoline in the presence of new car average efficiency standards may be different from the type desirable for forecasting the market share of station wagons for forecasting gasoline prices and consumption two months hence, a simple time-trend extrapolative model may be far superior to one including a detailed representation of the economic and engineering relationships. Conversely, for evaluating the impact of policy changes, such as oil decontrol on gasoline price and consumption, the time-trend extrapolative approach would be useless, while a structural modeling approach could be quite effective [11]. Realizing that different models are appropriate for different purposes, the EMF has attempted simply to delineate the capabilities and limitations of various models without undertaking the more difficult task of model evaluation.

Finally, potential working group members must be convinced of the value of the process to them as individuals. Each volunteers time and many contribute computer costs. If the expected rewards for participating are primarily public criticisms of their models, especially criticisms built on weak foundations, then the voluntary participation could be reduced notably. This may still become an issue. However, if the assessments are objective, and if they recognize strengths along with weaknesses, more telling comparisons probably are possible without discouraging the participation of the modelers. In this way, the Forum may include some elements resembling professional peer review but without the academic apparatus of manuscript refereeing.

The current movement is toward deeper model comparisons. At the same time, it is crucial that the issue focus not be lost. For the fifth study, the EMF staff devoted to model comparisons has been significantly expanded. Staff members currently are examining methodological differences among oil and gas supply models, to the extent possible, without hands-on experience. This examination was started even before the working group convened, with the expectation that the group will encourage and participate in this process. This expectation is being realized. The extent to which the EMF evolves towards deeper model comparisons and evaluations is an open issue. Its resolution depends upon such factors as the future EMF budget, the preferences of working group members, and the willingness of working group members to

expose their models to critical review.

A related issue is whether or not third-party model operation should be introduced into the EMF. In the current mode of operation, all model runs are made by the model developer, not by the EMF staff. Benefits of changing this mode of operation would be associated with the opportunity for a more scientific, objective, and complete examination of individual models. There would be several high costs. First would be the requirement for staff members to learn software that is model-specific. Second, there would be a weaker linkage between the evaluation process and the modelers, possibly resulting in less interchange of information. Finally, model assessment is far more costly than the current EMF procedure. Therefore, it is expected that little if any independent third-party operation will be introduced.

The role of the EMF in an academic institution such as Stanford University, where the educational progress of students is a key concern, raises another issue. In the past, Stanford students have participated in the EMF, but to a relatively limited extent. Currently, however, eight Stanford graduate students are participating directly in ongoing EMF studies or in preliminary issue identification and exploration for possible future studies. Many of these students are looking into methodological issues or are comparing methods, supporting data, or econometric techniques underlying participating models. This activity contributes simultaneously toward the academic goals of Stanford University and toward deeper, more telling model comparisons and evaluations. Thus, responsiveness to the educational goals seems at the same time to allow responsiveness to the call for increased critical evaluation, while maintaining the issue focus.

The final unresolved issue involves costs. Participants in an EMF study volunteer their time and generally computer expenses. Although most feel a sense of professional obligation and perceive a learning experience in participation, which justifies the donation of their time, the expense of running the sometimes quite costly models is not so easily absorbed. This policy has in the past excluded some potential participants who simply could not afford to participate without compensation. While the EMF has made an effort to help participants find a way to cover the computer costs of running the models, this policy may have, in the past, and probably will, in the future, exclude some of the models.

Many private and public sector organizations benefit extensively from the studies but bear none of the costs. Plans, therefore, are under way to raise money to cover the out-of-pocket expenses of study participants whose organizations cannot bear these costs. What kinds of costs should be covered, to what extent, and by whom are unresolved at present.

SUMMARY AND CONCLUSION

The Energy Modeling Forum, organized almost three years ago as an experiment to improve communication between energy decision makers and energy modelers, has been succeeding although many issues remain unresolved. Future success depends upon continuing cooperation from the broad community of energy modelers, planners, and policymakers. Constructive critiques, suggestions, and constant efforts to improve interaction will be important factors in fulfilling the objectives promised in the original design of the Energy Modeling Forum.

DISCUSSION

Dr. Greenberg (DOE): I am having a little trouble relating your description of the modeling forum activities with the subject of validation. I would like to take a couple of interpretations and have you interrupt me if I say something that is wrong. As I understand the overlap between modeling forum objectives and activities and the subject of validation, you have described a technique of model comparison. I don't fully accept your use of the term model comparison. All I see is a collection of forecasts coming from different models, published in a report. That to me means something different than model comparison.

Dr. Sweeney: I agree. If it were just a set of seven forecasts on what will be the energy supply and demand, published in a report, that would not be model comparison. There is a lot more in a typical study than seven forecasts.

First of all, a lot of the benefit of what goes on never appears in that piece of paper that gets published.

Dr. Greenberg: Why not?

Dr. Sweeney: A lot of what goes on is a learning experience associated with people having different perspectives learning about the usefulness of models. That is something that is hard, and virtually impossible to capture without participating in this entire experience.

We do put on paper much of the final reports. Now, let me ask you a question. How long was the study that you read when you responded to that? Roughly, how many pages did--30 pages or 700?

Dr. Greenberg: Well, it was far closer to 30.

Dr. Sweeney: Okay. The problem is you have only looked at the first report. We produce a report, and this is not your fault; it is because the second report is just going to the mail right now. We produce a 30-page report which is aimed at a broad group of people. It tries to focus more on the information that is generated from the process, about the results that are reported for policy purposes.

Then we have a second report which goes much more deeply into the model comparisons, and you can't do that in 30 pages. The background report for the coal and transition is 700 pages, which goes a lot more deeply into the model comparisons, documents the output of the models at several critical stages. I mean you can't just say what would be the quantity that would be produced under the various situations. We try to compare how each of the models behave at several critical points, or at least how many critical points we have examined in the models. It varies from model to model. I would anticipate, in the U. S. oil and gas supply function study, that we will look at estimates of drilling activity, reserve findings, production, reserves over time, reserves to production ratios, and so forth for oil and gas for each of the models under a set of different scenarios to see how each one of those variables differs in response to changes in input variables to the models.

Those are things that are conducted, and my feeling is that when you get to that stage you do have an excellent tool for examining the differences in behavior amongst the various models, and you stress the models in different ways. Certainly the 30-page report doesn't include everything hopefully the 700-page report does, and that is, again, for people to judge how well we have done that.

FOOTNOTES

REFERENCES

[1]   Energy Modeling Forum, "Aggregate Energy Elasticity Estimates,"
      Working Paper EMF 4.4, Stanford University, Stanford, California,
      September 1978.

[2]   Energy Modeling Forum, Coal In Transition: 1980-2000, EMF 2,
      Stanford University, Stanford, California, July 1978.

[3]   Energy Modeling Forum, "Electric Load Forecasting,"
      Working Paper EMF 3.11, Stanford University, Stanford,
      California, August 25, 1978.

[4]   Energy Modeling Forum, Energy and the Economy, EMF 1,
      Stanford University, Stanford, California, September
      1977.

[5]   Greenberger, M. "Closing the Circuit between Modelers
      and Decision Makers," EPRI Journal, Electric Power
      Research Institute, Palo Alto, California, October
      1977, No. 8, pp.6-13.

[6]   Hogan, W.W., "The Energy Modeling Forum: A Communication
      Bridge," paper presented at the 8th Triennial IFORS
      Conference, Toronto, Canada, June 20, 1978.

[7]   Hogan, W.W., "Energy Models: Building Understanding
      for Better Use," paper presented at the Second Lawrence
      Symposium on Systems and Decision Sciences, Berkeley,
      California, October 3-4, 1978.

[8]   Stanford Institute for Energy Studies, "Stanford-EPRI
      Workshop for Considering a Forum for the Analysis of
      Energy Options Through the Use of Models," Special
      Report EPRI EA-414-SR, Electric Power Research
      Institute, Palo Alto, California, May 1977.

[9]   Sweeney, J.L., "Energy and Economic Growth:
      A Conceptual Framework," Occasional Paper OP 3.0,
      Energy Modeling Forum, Stanford University, Stanford,
      California, November 1978. Also appearing
      in Symposium Papers: Energy Modeling and Net
      Energy Analysis (Colorado Springs, Colorado,
      August, 21-25, 1978), Institute of Gas
      Technology, Chicago, Illinois, 1978; and in Perlmutter,
      A.; Kadiroglu, O.K.; Scott, L.; eds.; Proceedings
      of the International Scientific Forum on an
      Acceptable World Energy Future, Ballinger
      Publishing Co., Cambridge, Massachusetts, 1979
      (forthcoming).

[10] Sweeney, J.L., and Flaherty, M.T., "Methodolgies
      for Petroleum Product Price Forecasting: A Review,"
      in Topics in Energy, Data Resources, Inc.,
      Lexington, Massachusetts, September 1978.

[11] Sweeney, J.L. and Weyant, J.P., "The Energy

Modeling Forum: Past, Present, and Future,"
Planning Paper EMF PP 6.1, Energy Modeling
Forum, Stanford University, Stanford, California,
February 1979. Also appearing in the Journal of
Business Administration (forthcoming).

ENERGY MODELING FORUM
SENIOR ADVISORY PANEL

Mr. Charles J. Hitch (Chairman)
President, Resources for the Future

Dr. Philip Abelson
Editor, Science

Dr. Harvey Brooks
Professor, Harvard University

Mr. David Cohen
President, Common Cause

Mr. Gordon R. Corey
Vice Chairman, Commonwealth Edison

Dr. Floyd L. Culler, Jr.
President, Electric Power Research Institute

Mr. Charles Di Bona
President, American Petroleum Institute

Mr. Herman M. Dieckamp
President, General Public Utilities Service Corporation

The Honorable John D. Dingell
Member, United States House of Representatives

The Honorable Joseph L. Fisher
Member, United States House of Representatives

The Honorable William P. Hobby
Lieutenant Governor of Texas

Mr. Jack K. Horton
Chairman, Southern California Edison Company

Mr. W.F. Kieschnick, Jr.
Vice Chairman, Atlantic Richfield Company

Dr. Henry R. Linden
President, Gas Research Institute

Mr. Guy W. Nichols
President, New England Electric System

Mr. John F. O'Leary
Deputy Secretary, United States Department of Energy

Dr. Alan Pasternack
Member, California Energy Commission

Dr. John Sawhill
President, New York University

Dr. Chauncey Starr
Vice Chairman, Electric Power Research Institute

The Honorable Morris K. Udall
Member, United States House of Representatives

Adam Borison
Research Assistant

John Lindsay Bower
Research Assistant

Patrick Coene
Research Assistant

Wendelin Dintersmith
Administrative Assistant

Darylin Druhe
Secretary

Mark Edgerton
Research Associate

Kathleen Favor
Research Assistant

Elizabeth Heck
Secretary

Charles Kolstad
Research Assistant

Douglas Logan
Research Assistant

Susan Missner
Research Associate

Zakia Rahman
Research Assistant

Dorothy Sheffield
Research Associate

Pamela Sherby
Secretary

Nancy Silvis
Editorial Assistant

James Sweeney
Executive Director

John Weyant
Associate Director

Thomas A. Wilson
Research Assistant

ELECTRIC LOAD FORECASTING:

PROBING THE ISSUES WITH MODELS

Bernard H. Cherry
Vice President
GPU Service Corp.

rking Group Three of the Energy Modeling Forum was convened in
cember of 1977 in order to evaluate the approaches utilized for the
recasting of electric energy demand in several utilities across the
untry. The final report of this activity was published in the
ring of 1979.

ter the selection of load forecasting as a topic for the third working
oup, potential participants were identified. There was an attempt
 get a cross section of participants in the utility sector in terms
 the type of service area, geographical location, and the economy of
e service area. There was no attempt to pick the best ten electric
ility load forecasting approaches, but rather to randomly select a
mber of utilities across the country. Invitations were issued to
out 25 utilities. Invitations were also forwarded to university
delers active in energy modeling and finally to a number of state
d federal regulators active in the field. The ten models which were
timately used in the load forecasting study are shown in Figure 1.

e initial meeting of the group was aimed at trying to define a
mber of issues: (1) what to evaluate in the study; (2) how to com-
re the results of these evaluations; and (3) how to select the
enarios for evaluation. There was substantial concern about the
timate use of a comparison of various utility and university
dels. There was concern particularly among the utility partici-
nts of the perception of study results by one or more of their
spective regulatory agencies. While this might be viewed as an
necessary concern, those who are interested in the validation and
mparison of models will have to come to grips with this very real
ncern.

e Working Group meetings (the first two in particular) resembled,
sort of encounter session. There were very strong views and opinions
ld by a large number of the participants in the sessions, and the
solutions and accommodation of some rather divergent points of view,
came a very difficult and tedious task for all of us. However, it
 to the credit of all of the members of the Working Group, that the
y issues were resolved. The difficulties which were faced in the
rly meetings can be highlighted by the problems associated with
e selection of a title for the study: Electric Load Forecasting:
obing The Issues With Models. The word "Probing", in particular,
 one which required a good deal of negotiation to gain ultimate
reement. There was a resistance in using any words which conveyed
ore quantitative comparison than "Probing". Rejected were words

like evaluating, comparing, and investigating.

The next issue was to decide exactly how the evaluation could go forward. The key was the definition of a base or reference case for each modeler. It was ultimately agreed that the most efficient approach was to use each modeler's current planning scenario as a base case with certain adjustments relating to time of day pricing and appliance efficiency standard effects removed. This provided a relatively quick opportunity to run calculations based upon input which already included the best judgements of all of the modelers. In the course of the discussions, it was agreed that it would be extremely time consuming to start from scratch and construct an artificial set of data, and have each of the participating modelers attempt to run some standard computation. While it was recognized that there were deficiencies in the approach chosen, particularly relating to an explicit understanding of the judgements which went into the base data, the participants agreed that this was the most efficient way of proceeding.

The selection of the scenarios to be evaluated beyond the reference was the subject of an entire meeting and after discussion of a large number of issues which could be important in the future, about a dozen were selected as being worthy of future evaluation. Of those, seven were viewed as being able to be evaluated by some of the models which were being investigated.

The scenarios which were ultimately evaluated in the forecasting study are shown in Figure 2.

Figure 3 indicates that while at least one model was able to evaluate every scenario, not all models were able to evaluate all scenarios. It should be noted in Figure 3 that some modelers did not evaluate all scenarios due to personnel availability and time constraints.

Two major objectives were met in the EMF 3 Working Group. First, the experiments identified and illuminated key forecasting issues, and second, the interactions among the model developers and users improved the understanding of the model's capabilities and limitations. The second objective, in the author's view, was most important in this evaluation, inasmuch as there was little formal communication between model developers and users in various utilities prior to the convening of this working group.

The major findings of the analysis and the discussions of the group are shown below:

.   Future electricity consumption, given the assumptions
    of the modelers, is forecast to grow more slowly than
    in the past as shown in Figure 4.

.   Load shape, as well as peak demand and electricity con-
    sumptions, is a critical determinant of future genera-
    tion capacity requirements. Implementation of time-

of-use pricing and load management techniques will make forecasting load shape even more crucial in the future.

. Increases in real electricity prices significantly reduce the consumption projected using models which explicitly include prices, although the degree of response was substantially different among the models. (See Figure 5). As electricity prices change, capturing this effect in the forecasts will continue to be important.

. Increases in the relative prices of other energy sources, e.g., natural gas, cause increases in the electricity consumption projected using models which explicitly include the prices of these fuels. As prices of competing fuels change, it will be vital to capture these effects in the models.

. Combined historical data from many regions represent a largely untapped source of information, but the appropriate use of these data was hotly debated in the group. The significance of the issue has been highlighted by the observation that the price impacts are much larger in the few models estimated with combined data than in the models based on data from a single utility area. One view is that the empirical estimates obtained using single utility area data are most relevant, while the other view is that the estimates derived from combined data are most appropriate.

. Adoption of efficiency standards for appliances and construction can significantly influence projections of electricity consumption. Thus it is important to incorporate in the models the effects of standards and other regulatory changes.

It is interesting to note that some of the conclusions which resulted from the analysis did not result from direct output of the various computations, but rather resulted from interaction between the various participants in the sessions. These derived results turn out to be more important than some of the results which are direct output.

It should be pointed out that the conclusions were subject to drafting by the entire Working Group and, because of that, perhaps emerged somewhat less penetrating than if the drafting by committee technique was not used. This is a symptom which is probably unavoidable and it is for this reason that the real benefit from these kinds of efforts are derived by the participants themselves. Normally only a very small fraction of the benefit of studies of this type is conveyed in published reports.

Two of the key results which were seen in the evaluation related to future electricity consumption and the effect of changes in real-electricity prices. In the first case, the future electricity consumption --- the group found that the mean projected compound annual growth rate for all of these modelers was about 4.7%, substantially lower than the historical trend of about 7%. This reflects, over a wide range of service areas, a consensus that future consumption is likely to grow at a much slower rate than in the past.

The second key result was that there was a large projected variation in the reaction of energy demand to changes in real electricity price. This is seen in the contrast between those models which were more regional or national in scope where a much stronger response to changes in energy price is predicted and those models which are more service area specific and predict a lower elasticity. This result led to a good deal of discussion. A long and controversial discussion resulted as to which was the right approach. The Working Group was unable to arrive at an agreement as to which was the right approach, but recommended that this was an area where substantial additional work could be done.

The recommendations which the Working Group agreed upon are shown below:

.   Uncertainty of inputs and models suggests that a comprehensive forecasting effort should provide a range of forecasts.

.   Econometric and engineering models should be used together, either through the use of complementary models, or in a single model.

.   Current data collection procedures should be scrutinized to improve their cost effectiveness for forecasting purposes.

.   New data collection efforts should be undertaken.

.   Forecasting methodologies should be improved in concert with the data cases.

.   Programs aimed at defining evaluation criteria and procedures should be examined to insure that forecast model development is not stifled.

.   Cooperation among utilities with similar forecasting problems should be encouraged.

Specifically relevant to the subject of evaluation and comparison of models, there was a very strong feeling in the Working Group that programs aimed at defining evaluation criteria and techniques should be carefully examined to assure that the model development was not stifled. There was a strong belief that the "capitalistic

pproach" in modeling has worked very well in the past and is likely
o continue to work in the future. It was agreed that any programs
hich tend to impede the free development of new approaches and ideas
ere likely to be counterproductive.

nother recommendation of the group was that cooperation among the
tilities with similar forecasting problems should be strongly encour-
ged. There have been some follow-on efforts associated with this
ctivity. A group has been convened under the auspices of EPRI called
he Utility Modeling Forum (UMF). This group has a focus similar to
he Energy Modeling Forum but more narrowly aimed at utility modeling
roblems.

n conclusion, some observations which I think are relevant to the
ocus of the workshop should be emphasized. The key benefit of
odeling inter-comparison should be to gain a better understanding
f the issues which are important in modeling, the structure of the
odels and the ability to deal with those issues. In the forecasting
rea (this would probably be true in any controversial analysis
ffort), the major benefits and many of the key findings which result
rom the effort, are not communicated in the written materials or
eports, and emerge only in the group meetings. Therefore, the
ajor benefits go to the active participants. This is a fundamental
ngredient of this sort of program and would argue against any kind
f independent third party evaluation of models.

inally, there is a major need for continued communication between
he various modelers, particularly in the electric energy forecasting
ector. This will ultimately lead to a better understanding of fore-
asting problems.

FIGURE 1

MODELS USED IN THE ELECTRIC LOAD FORECASTING STUDY

1. Commonwealth Edison Company, Econometric Model (Comm. Ed.)

2. Oak Ridge National Laboratory, Residential Energy Demand Model (ORNL-REDM)

3. Oak Ridge National Laboratory, State-Level Electricity Demand Forecasting Model (ORNL-SLED)

4. Tennessee Valley Authority, Load Forecasting Model (TVA)

5. Consumers' Power Company, kWh Sales Model (CPC)

6. Florida Power and Light, Simulation Model (FPL)

7. Northeast Utilities, Electric Energy Demand Forecasting Model (NU)

8. University of Texas, Baughman-Joskow Regionalized Electricity Model (Baughman-Joskow)

9. Wisconsin Electric Power Company (WEPCO)

10. General Public Utilities (GPU)

FIGURE 2
DEFINITION OF SCENARIOS USED IN THE ELECTRIC LOAD FORECASTING STUDY

Scenario	**Definition**
1. Reference Case	Each modeler's current planning or "base" scenario with time-of-day pricing and appliance efficiency standard effects removed
2. Price of Electricity--1 (Average Price Increase)	10% increase in the average price of electricity over the price used in the reference case
3. Price of Electricity--2 (Demand Charge Increase)	10% increase over the reference case in any demand charge with energy charges held at the reference case values
4. Price of Electricity--3 (Energy Charge Increase)	10% increase over the reference case in energy charges with demand charges held at reference case values
5. Competing Fuels Price	20% increase over the reference case in the delivered prices of oil and natural gas
6. Appliance Efficiency Standards	Efficiency improvements relative to typical equipment or buildings in place in 1974 were specified for major appliances and residential buildings
7. Technological Change-- Cogeneration	10% incremental tax credit on the cost of investment for cogeneration facilities
8. Time-of-Day Pricing	6 to 1 price ratio for on-peak to off-peak consumption for all customers with the peak period being 8:00 a.m. to 8:00 p.m. on non-holiday weekdays

FIGURE 3

MODELS USED TO ADDRESS THE SCENARIOS

		Scenario							
		1	2	3	4	5	6	7	8
	Model								
1.	Comm. Ed.	X	X	X	X	X			
2.	ORNL-REDM	X	X			X	X		
3.	ORNL-SLED	X	X			X			
4.	TVA	X	X	X	X	X	X	X	
5.	CPC	X	X			X	X	X	
6.	FPL	X					X		
7.	NU	X					X		X
8.	Baughman-Joskow	X			X				
9.	WEPCO	X			X				
10.	GPU	X							

FIGURE 4

Reference Case Electricity Consumption

TOTAL ENERGY USE
(Kilowatthours, 1975 = 100)

0: GPU	5: CPC
1: Comm. Ed.	6: FPL
2: ORNL-REDM	7: NU
3: ORNL-SLED	8: Baughman-Joskow
4: TVA	9: WEPCO

FIGURE 5

Change in Electricity Consumption with 10% Price Increase, Relative to Reference Case
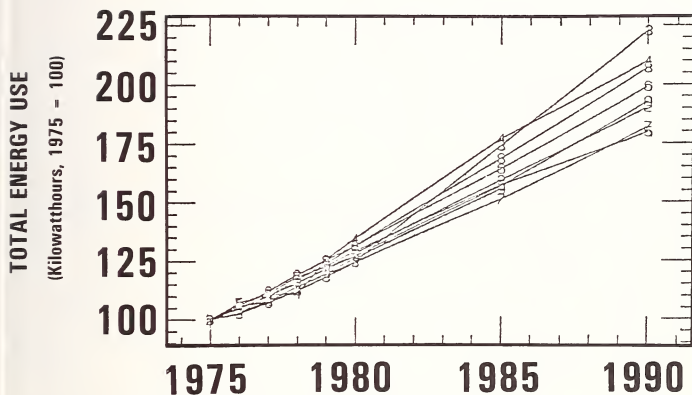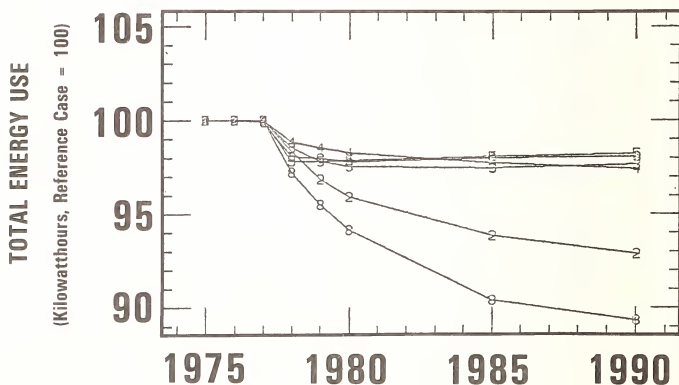
0: GPU
1: Comm. Ed.
2: ORNL-REDM
3: ORNL-SLED
4: TVA

5: CPC
6: FPL
7: NU
8: Baughman-Joskow
9: WEPCO

Mr. Cherry: Let me make one quick point that had been overlooked. The major benefit in this activity--I don't know if you want to call it model validation or model probing or model investigation--is to the participants, to the people who actually plan the cases. I think we are looking here at controversial kinds of questions.

Maybe only 10 percent of the benefit and the knowledge which came out of the nine months of work that we did eventually appears in the published material. I think that includes all of the published material. I think that there is a very large leverage to be involved in the detailed inter-actions which went on in the various sessions. I think those who did that derived a major benefit from that activity.

Dr. Wagner (U. of NC): Actually, your last comment was a partial answer to my question. I was just wondering if you could indicate what happened in the utility companies as a result of this. Did they build some more models or revise the ones they had? Just what did occur?

Mr. Cherry: Well, I can tell you what happened in my company, specif-ically, in forecasting one of the things I am responsible for. We were able to, in some greater detail than in the past, understand what other people were doing. These were the people who had problems similar to ours.

I think this allowed us to save a lot of time and energy in choosing correct paths, perhaps more efficient paths for improving our models. We did select a number of improvement opportunities for change, and this made that whole process easier. I think a number of others did the same thing.

Dr. Nissen: If the object of validation and assessment procedures is to make the understanding of models public and explicit and understood, then part of the social capital of the analysis business must also be explained. Much of what we are trying to understand here could produce things which would have been valuable to you. Apparently, you found it very difficult to communicate to the world outside, because you understood it so late, or it was so controversial until the understanding came amongst the models. Therefore, the coming to understand it was most of the work that you should get done.

The communicating of it, in a way which would be recognizable, is left for another generation of forums, I guess. What would assessment have done, and what would you have used out of it, in this process, if this kind of a meeting is successful and another forum was held three years from now?

Mr. Cherry: I have two reactions to that question. One, it is a
hard question. I am not sure. I think, perhaps, the result of this kind
of activity could make it easier by conditioning the people in the field
to be a little more forthcoming with results. I think that, perhaps,
may be one of the major benefits.

A second benefit is found in terms of comparison of results of work
criteria and computations. We ask, where do you start--with basic data
or results? I would hope that during the next two years that those types
of questions would be "ventilated," as Bill Hogan says. I hope that there
will be more of a consensus on what the right process is, and that is
something that we had a lot of trouble coming to grips with in our little
activity.

I had hoped--and we all recognized when we did it--that the result was
something of an expedient. It was constrained by the time and resources
that we had available, and that was the best thing to do at the time. I
hope that these kinds of future activities would give some better thought
and direction to that kind of a process.

ASSESSING THE ICF COAL AND ELECTRIC UTILITIES MODEL

Neil L. Goldman and James Gruhl

M.I.T. Energy Laboratory, Model Assessment Group

I. Objectives and Types of Model Assessment Activities

Because there is such a variety of ways and means for evaluating a model, the first step in an assessment process should be the development of a strategy. In fact, such a strategy must be very carefully chosen and orchestrated to make proper use of the inevitable limitations in time, funds, and manpower available for the assessment. In order to make choices between alternative assessment paths, the objectives or goals of the assessment must first be clearly understood. Some possible objectives include:

    (1)   validate specific past applications,

    (2)   validate generic future applications,

    (3)   suggest model improvements,

    (4)   create a resource group having expertise on the model,

    (5)   establish credibility among users,

    (6)   test transferability or usability by others, and

    (7)   further the state of the art of model assessment.

The most obvious, and perhaps best defined, of these objectives is the validation of the input data and structural form of the model relative to specific applications. It would probably be more useful to make statements about the appropriateness of a model for contributing information to future policy decisions in generic application areas. Two objectives that would be difficult to achieve simultaneously would be (1) suggestions for model improvements, and (2) establishing model credibility. The first of these suggests a series of model versions, while the second suggests a single model version established at the beginning of the assessment.

Once the objectives have been decided, there are a number of alternative settings and depths for the assessment process. Some of these alternatives result from the possibility of different assessor identities. For example, the assessors could be any of the following:

    (1)   model builder,

    (2)   model sponsor,

    (3)   independent third-party,

    (4)   several model builders, such as a forum, or

    (5)   several third-party assessors.

In addition, the assessors could address either a single model or several comparable models. There could be very different expectations from the assessment process depending upon this choice of setting. For instance, the model builder could obviously provide a very cost-effective assessment, but credibility would be difficult to establish under such circumstances.

For each possible assessment setting outlined above there are four potential depths:

(1) <u>literature review</u>: survey and critical analysis of the published literature on the structure, implementation, and applications of one or more models;

(2) <u>overview</u>: literature review plus analysis of computer codes and other unpublished materials;

(3) <u>independent audit</u>: overview plus the conduct of model exercises designed by the assessor and executed by the modeler with the assessor "looking over the shoulder;"

(4) <u>in-depth</u>: independent detailed assessment of model formulation, structure, and implementation with the assessor in control of the model and associated data base.

The most cost-effective of these depths will depend upon a number of model characteristics, most particularly, model maturity. If a model is very mature, it is probably worthwhile to go to the expense of an in-depth assessment. If it is immature, then an audit or overview might be sufficient for reaching major conclusions. Size, structure, complexity, execution costs, previous applications, and previous assessments are all aspects that should contribute to the decision on the most cost-effective depth. It might be noted that the classical validation process has consisted of in-depth assessment by model builders, audit roles for model sponsors, and literature review or peer review by independent parties.

As has been pointed out by Saul Gass in several of his papers, an important way to limit the assessment process is to limit its scope (see Table 1). First, decisions must be made concerning the version(s) of the model that is to be assessed, and the types of model applications at which the assessment process is to be aimed. Point 2 of Table 1 defines different aspects of the model that can be evaluated in an assessment: documentation, validity, verification, or operational characteristics.

The ability to assess model documentation adequately will depend to a large degree upon the content and amount of written material that has been produced by the model builders. There are a number of different items that must be included in the documentation:

(1) Methodology and Philosophy Behind Model Structure - mathematical formulation, parameter estimation, and computer code,

(2) Data - description of data that have been used, preparation of new inputs and parametric data,

(3) Use of Code,

(4) Past Uses and Results,

(5) Range of Applicability of Model, and

(6) Descriptions of All Validations Performed - by model builder or by independent parties.

Table 1

DIFFERENT CATEGORIES REPRESENTING VARIOUS SCOPES OF ASSESSMENTS

1. Specific Applications of Interest

    1.1 Validation in context of specific applications, ranges of variables, degree of aggregation required, absolute values versus policy perturbation studies

    1.2 No specific cases, just an assessment that provides the foundation for generally evaluating model accuracy

2. Aspects to be Assessed

    2.1 Documentation - of structure, validations performed, parameter estimation, past uses, applicability, computer code and use

    2.2 Validity - informed judgment about model logic and empirical content

        2.2.1 Data or input validity - empirical implementation, quality of updating procedure

        2.2.2 Logical or structural validity

        2.2.3 Predictive or output validity

    2.3 Verification - accuracy of algorithmic and computational implementation

    2.4 Operational characteristics - flexibility, extensibility, transferability (training required, ease of use, modeler independence from model and model knowledge), time and cost efficiencies.

Before discussing validation and verification techniques it is necessary to define some terminology. The model is considered to be built from historical or other data observations. The inputs are defined as any values that change for different applications of the model. The parameters and structural elements are those aspects of the model that are meant to stay the same for different sets of model runs. With these definitions in mind, Table 2 illustrates different types of validation and verification techniques that have been found described in the literature or have been postulated by us. These validation techniques are essentially two-part processes. The first part involves examinations or actions that are performed on parts of the model. The second part of the process involves an assessment of the validity of those actions as measured by any of the seven bases for comparison listed at the end of Table 2.

In previous documents we have discussed several of these validation techniques, so a lengthy discussion would not be appropriate here. We have chosen point 5.4 in Table 2 as a means of summing up our discussion of validation techniques. In many ways this point represents the ideal final result of an assessment, that is, a probabilistic measure of the output validity of the model. The class of models for which this ideal measure can be developed has not been clearly established. It is likely that this ideal probabilistic measure can only be bounded from above and below on the basis of simplified assumptions and techniques, possibly including linear or nonlinear analytic representations of the model's input-output response surface. For simple enough representations of the model it might be possible, either analytically or through the use of Monte Carlo techniques, to propagate input uncertainties through structural uncertainties to create measures of output uncertainties. Besides being difficult conceptually, the process of developing quantitative measures of predictive quality is likely to be hampered by:

(1) the fact that it may be as time-consuming a process as the whole model building procedure,

(2) it will be application specific,

(3) funding requirements are generally not appreciated by sponsors, and

(4) decision makers are not now insisting on such displays of predictive quality.

The final aspect of the assessment scope is an evaluation of the model's operational characteristics. These characteristics can generally be categorized as:

(1) Ease of Updating Data - different types of applications, changes of levels of aggregation,

(2) Flexibility through Input and Parameter Changes - different applications made possible through changes only in inputs and parameters,

112

Table 2

VALIDATION AND VERIFICATION TECHNIQUES

ACTIONS: EXAMINATIONS OR CHANGES

OBSERVED DATA:

1.1 Examinations of the observed, historical, or estimation data

OBSERVATIONS-TO-STRUCTURAL:

2.1 Observed data perturbation effects on structure and parameters
2.2 Propagation of estimation error on structure and parameters
2.3 Measure of fit of structure and parameters to observed data
2.4 Effect of correlated or irrelevant observed or estimation
     data on structure and parameters
2.5 Sensitivity analysis: quality of fit of structure and
     parameters to observed data for altered structure and
     parameters (includes ridge regression)

OBSERVATION-TO-INPUT:

3.1 Effects of correlated or irrelevant observed data on outputs

INPUT:

4.1 Base case or recommended input data examinations

INPUT-TO-OUTPUT:

5.1 Examine outputs with respect to base case input data
5.2 Simplify, e.g. linearize, this relationship to provide
     understanding
5.3 Simplify (e.g. linearize) structural form analytically, or
     group parameters to provide better understanding,
     elimination of small effects, grouping of equations,
     grouping of parallel tasks
5.4 Develop confidence measure on outputs by propagating input
     error distributions through structural error distributions

STRUCTURE:

6.1 Structural form and parameter examinations
6.2 Respecification, that is, make more sophisticated some
     of the structural components
6.3 Decompose structure physically or graphically
6.4 Provide new model components to check effects of assumed
     data or relationships

Table 2 (continued)

STRUCTURAL-TO-OUTPUT:

7.1 Examination of outputs for various structural and parametric
     perturbations and error distributions

OUTPUT:

8.1 Examination of outputs

OUTPUT-TO-INPUT:

9.1 Examination of optimal inputs (controls) to reach target outputs
9.2 Contribution analysis, percentage changes in outputs due to
     changes in inputs

BASES FOR COMPARISON

A. Comparison with other empirical models
B. Comparison with theoretical or analytical models
C. Comparison with hand calculations or reprogrammed versions
    of model components
D. Data splitting on observed data, by time or region
E. Obtain new estimation/prediction data with time, new
    experiments, or in simulated environments
F. Examination of reasonableness and accuracy, that is,
    comparison with understanding
G. Examination of appropriateness and detail

(3) Extensibility of Structure - ease of structural changes for new applications,

(4) Interpretability of Output,

(5) Efficiencies - in time and cost,

(6) Understandability - transparency of modeling philosophy and structural relationships, and

(7) Transferability - accessibility of documentation, training required, ease of use by others including the amount of art versus science necessary in the operation of the model.

The final phase of the entire assessment process should be an evaluation of the assessment itself. There are several techniques that can be used for this evaluation including:

(1) Model Builder's Critique - identification of assessor misconceptions, value to builder of validation, points of contention, completeness of coverage of new errors builder has found since turning model over to assessors,

(2) Complete Coverage of Model's Components - balance or evenness of assessment efforts, discussion of skill limitations of assessment team,

(3) Complete Use of Available Validation Techniques - numbers of actions and bases for comparison used,

(4) Number of Inexplicable Results - problem areas the assessors were not able to fully unravel,

(5) Comparisons with Other Validations - possibly of the same model at the same time, and

(6) Test of Time - evaluate validation conclusions on model accuracy using new historical data.

This concludes the general discussion section of this paper. The remainder of the paper concerns the M.I.T. Energy Laboratory's assessment of the ICF Coal and Electric Utilities Model (CEUM). Because this assessment is only partially completed, with the entire story, including the model builder's rebuttal, not yet available, the scope of material that is appropriate to discuss at this time is somewhat limited. In the following sections we therefore present: a brief history of the model, an assessment of the model documentation, a description of the model structure, a discussion of the model's linear programming matrix and objective function, a discussion of important structural issues, and finally some general ideas concerning the overall design of the CEUM and strategies for further assessment.

II. History and Applications of the ICF Coal and Electric Utilities Model

The evolution of ICF's Coal and Electric Utilities Model (CEUM) is somewhat complex. The first output of ICF's coal modeling work was the PIES Coal Supply Analysis, (1), developed in 1976 for the Federal Energy Administration (FEA) report, National Energy Outlook (a 1976 update of the Project Independence Report). This coal supply methodology was reviewed by Resources for the Future, (2), and extensively evaluated and critiqued in an Electric Power Research Institute (EPRI) report by Richard L. Gordon of Pennsylvania State University (3). As a major extension of its coal supply analysis for PIES, ICF developed the National Coal Model (NCM), (4), for FEA in 1976. The NCM was also critiqued in Gordon's EPRI report, (3).

The Coal and Electric Utilities Model Documentation, (5), of July 1977, is essentially a retitled and enlarged version of ICF's 1976 description and documentation of the NCM, (4). The enlargement consists simply of adding an additional appendix to the NCM report containing a series of memorandums, written over a period of about a year to mid-1977, on possible changes and refinements in the model. We note here that the CEUM Documentation is the subject of the assessment in Reference (10).

The CEUM was developed by ICF as an energy policy planning tool. It was designed to address policy and planning issues related to the coal and electric utility industries and can be used to analyze:

o   regional coal production and consumption

o   regional coal prices

o   coal transportation requirements

o   utility capacity requirements

o   utility fuel use

o   impacts of changes in oil prices, planned generating capacity additions, and the growth rate of electricity consumption

o   impacts of government policies concerning:

- Clean Air Act Amendments

- western coal development

- regulation of strip mining reclamation

- Energy Supply and Environmental Coordination Act conversion orders

- taxes on oil and gas use.

Since 1977 several U.S. government agencies and EPRI have secured studies using the CEUM. The CEUM was also one of the coal models examined in a 1978 study conducted by Stanford's Energy Modeling Forum, entitled Coal in Transition: 1980-2000, (6).

116

Three ICF reports have recently appeared devoted primarily to presenting a series of case studies using the model. The first report, prepared for the Environmental Protection Agency (EPA), consists of two phases of an analysis dealing with the impacts of alternative new source performance standards (ANSPS), i.e., alternative changes in sulfur oxide emission standards. This study was undertaken to assist EPA in reviewing the current new source performance standard (NSPS) following the 1977 Amendments to the Clean Air Act. These amendments mandate the use, in new large fossil-fuel burning installations, of the best available technologies for pollution control. The two phases of the work involved two separate sets of scenario specifications on the meaning and costs of ANSPS. Both phases employed the model largely in the form reported in the CEUM Documentation, (5), with the entire data base updated. However, two major changes were made. First, partial scrubbing was allowed. Second, the target-year runs were made in a sequence such that information from earlier year runs could be used in later year runs, i.e., intertemporal constraints were incorporated. Previously, each target-year's solution was derived independently of those for other target years. The first-phase work was completed in late 1977 and the second phase in April 1978, but the documentation of the complete study was not reported until September 1978, (7).

A second report by ICF, prepared for the Departments of Interior and Energy (DOI/DOE), deals with the demand for western coal and its sensitivity to key uncertainties, and considers the question of the need for additional leasing of federal lands in the west. Again, some structural changes were made in the CEUM, and radically different basic demand assumptions proposed by the two agencies were employed. ICF's full report on this study was issued in June 1978, (8).

Finally, a third ICF report, prepared for EPA and DOE, again dealing with the impacts of ANSPS, was completed in September 1978, (9). This study involved still further revisions in the basic CEUM, utilizes demand assumptions closer to those used in the DOI/DOE study than to those in the earlier EPA study, and considers still another set of scenario specifications on the meaning and costs of ANSPS. It is suggested by ICF that the set of forecasts produced in this latest study should be given substantially more credibility than forecasts in previous studies because the CEUM is more refined, the scenario specifications employed are more up-to-date, and better estimates of scrubber costs are utilized.

A summary of the history and major applications of the CEUM is presented in Table 3.

Table 3

HISTORY AND MAJOR APPLICATIONS OF THE CEUM

January 1976 - May 1976	PIES Coal Supply Analysis
August 1976	RFF Evaluation of PIES Coal Supply Methodology
October 1976	National Coal Model (NCM) Documentation
July 1977	Gordon's Critique of NCM
July 1977	CEUM Documentation
July 1978	Energy Modeling Forum Study - Coal in Transition: 1980-2000
September 1977 - April 1978	CEUM EPA Study
April 1978 - June 1978	CEUM DOI/DOE Study
April 1978 - September 1978	CEUM EPA/DOE Study

III. Model Documentation

There are three tangible aspects of a policy model: the documentation, the implementation (computer code), and the experience and know-how of the model-builders, operators, and analysts who use and maintain the model. All three aspects are indispensable, but like the nucleus of a living cell, the model documentation should ideally contain all of the information from which the entire model organism can be regenerated. Model documentation, again ideally, should be oriented in each of three different directions: toward the user, toward the operator, and toward the analyst. User-oriented documentation, to be exhaustive, should contain the following information:

o   The motivation and objectives underlying the model development.

o   A description of the capabilities of the model and the scope of its applications.

o   A general explanation of the structure of the model.

o   A clear statement of all assumptions and restrictions imposed on the model.

o   A general description of all data inputs used and an explanation of their sources or derivation.

o   A description of all input parameters required by the model.

o   Instructions for the interpretation of model output.

o   A discussion of the costliness of using the model.

The model operator is the person (or persons) who provides the interface between the policy-oriented user and the computerized algorithms of the model. Documentation oriented towards the operator should ideally contain:

o   Detailed instructions for running the model with input and output formats and all model options fully explained.

o   A thorough explanation of the internal structure and logic of the computer program sufficient to enable the operator to modify the program where necessary for the particular needs of the user.

The model analyst is the person (or persons) who has an in-depth understanding of various phases of the model. The analyst may act as a resource for the user or may be an independent model assessor. Of course, all model-building groups include such analysts. The analyst may be needed to evaluate the usefulness of the model for various applications, the model's limitations, the reliability of the results, the appropriateness of the data, and other aspects of model performance. In addition, the analyst may contribute to complex modifications of the model or entire reconstructions. A policy model should not be

a package presented to the user as a black box; rather it should offer the user a general framework and specific tools with which he can produce a variety of output projections. The user must be able to shape the model and apply it in face of unforeseen problems and questions. It is in this task that the knowledge of the analyst is crucial.

In addition to the documentation requirements mentioned above, the analyst would find it useful to have documentation containing a complete and detailed description of the construction of the model. For the analyst, the documentation cannot be considered truly complete unless it is sufficient to permit replication of all aspects of the model. The analyst cannot work successfully with information that exists only in the minds of the model builders. This information must appear on paper, in an organized and readable form, before it can be part of an assessment. More specifically, the model analyst should ideally have access to:

o   A complete technical description of the model processes, including a precise and well-defined formulation of all analytical techniques used.

o   A complete description of all data inputs used, with sources or derivations specified.

o   A listing of the complete computer code, fully annotated with "comments" and summary explanations.

As a model is modified, addenda can be added to both the documentation and the computer code. Periodically however, such modifications should be integrated into the documentation and the computer code, in such a way that each becomes a rationalized whole. We believe that all aspects of documentation specified above should, to the extent possible, be a requirement in contracts for model development.

The CEUM Documentation, (5), is inadequate when measured solely against the above standards. The reader should note that many of the shortcomings in the model documentation have their origins in the dual role which ICF has played in the CEUM, being responsible for both model development and application. One consequence has been that emphasis is placed upon study documentation for clients with some model descriptions to motivate the analysis, and with careful attention to reporting key model input data, and in interpreting model output. With regard to this study documentation, ICF receives very high marks. To date, study clients have not required the other kinds of model documentation which we describe above. As a consequence that information is available only incidentally in the current documentation, and within the ICF modeling group. ICF has been candid in discussing model specification issues, and takes the view that formal documentation of these issues is not important to study clients providing they do not intend to execute the model independently. We only partially agree with this view. First, study users do require the model documentation as a reference for interpreting and analyzing study results. Secondly, potential model users and analysts require such documentation as the basis for evaluating the model approach, specification, and embodied research results. Finally, such documentation is a necessary condition for good scientific practice. We believe that ICF and their current clients have somewhat underestimated the importance of these reasons for augmenting the current study reports with model documentation.

With this background we now return to an evaluation of the CEUM model documentation using the guidelines suggested above. The introductory sections of the documentation include a discussion of the CEUM's capabilities and key characteristics but do not include a discussion of the motivation and objectives underlying the model development. This material is oriented toward a study client rather than a model user or analyst. The explanation of the model structure in Chapter 2 of the documentation is on a level that would ordinarily be sufficient for the user but not for the analyst. Unfortunately, even on this level of generality, the explanation in parts is highly misleading and gives little if any indication as to the true nature of the CEUM's structure. In particular, the so-called "non-technical flowcharts," supposedly illustrating the model's logic, create the impression that the model structure is in the form of a sequential decision process when in actuality it is a simultaneous process of constrained minimization. While ICF cautions in the documentation that these flowcharts are neither complete nor technically precise, the impression is created that the flowcharts present an accurate general picture of the model structure.

The data inputs used by the CEUM are extensively displayed in Chapter 3 of the documentation though in many places their derivations are only partially discussed. For example, the important concept of minimum acceptable real annuity coal prices is not adequately described. Our own description of this concept appears in Section 3.2 of Reference (10). In general, for each model component, the documentation does identify the data requirements and cites all sources of information.

Instructions for the interpretation of model output are not given in the documentation nor is there any discussion of the costliness of using the model. Furthermore, it is clear that the CEUM documentation was not written in a manner facilitating the model's transferability, and obviously as a commercial organization ICF may have very sound reasons for avoiding such explicitness.

A major accomplishment of our overview assessment effort has been the development of a complete, detailed, and well-defined mathematical formulation of the basic set of equations employed in the CEUM (see Reference (10), Chapter 2). An illustrative linear programming (LP) matrix displaying the basic structure of the model for one supply region and one demand region has also been developed and described in Reference (10), (for a description see Section V of this paper). This matrix is loosely based on a sample LP matrix, appearing in Appendix A of the CEUM documentation, which is incomplete, unclear, and lacking in adequate explanation. We believe that our own development of the CEUM's mathematical formalism represents a valuable addition to the model's documentation, and furthermore, is necessary if one is to fully understand how the model works.

In conclusion, we believe that the CEUM documentation does not provide interested groups the information required to evaluate, use, operate, or modify the model without the assistance of ICF personnel.

IV. Structure of the CEUM

The general structure of the CEUM, (5), consists of a supply component that provides coal, via a transportation network, to satisfy, at minimum cost, demands from both utility and non-utility users. The CEUM is static and regional. It generates an equilibrium solution through a conceptually straightforward linear programming formulation that balances supply and demand requirements for each coal type for each region. The objective function of the linear program minimizes, over all regions, the total costs of electricity delivered by utilities and the costs of coal consumed by the non-utility sectors. Regional levels of electricity generation and non-utility coal use are preset. The model shows how best to meet these exogenously determined final demands. The output of the model includes projections of coal production, consumption, and price by region, by consuming sector, and by coal type for the target year under consideration. The impacts of environmental standards for electricity generation from coal are also considered explicitly.

Table 4 outlines the basic elements of each of the four major components of the CEUM:

> (1)   Coal Supply
>
> (2)   Utility Demand
>
> (3)   Non-Utility Demand
>
> (4)   Transportation

Some key characteristics of the CEUM's major components are as follows:

o   Coal supply is disaggregated into 30 supply regions.

o   The model has the capability for considering up to 40 different coal types representing all possible combinations of five BTU content groups and eight sulfur levels.

o   The utility demand for steam coal is disaggregated into 35 demand regions.

o   Non-utility coal demand, exogenously specified by region, is disaggregated into 5 consuming sectors: metallurgical, industrial, residential-commercial, synthetics, and exports.

o   The electric utility demand for coal is determined endogenously by taking account of the exogenously specified total electricity demand by region and inter fuel substitution possibilities.

o   Transportation costs are based on rail and barge shipment rates.

o   Environmental standards for electricity generation from coal are considered explicitly through endogenous options to meet utility demands by use of coal types having appropriate sulfur characteristics and corresponding desulfurization costs.

Table 4

MAJOR COMPONENTS OF THE CEUM

SUPPLY	UTILITY DEMAND
. 30 Regions	. 35 Regions
. 40 Coal types	. 19 Coal piles
- 5 Btu categories	- 3 Ranks of coal
- 8 sulfur levels	- 6 Sulfur categories
. Existing capacity	- Metallurgical pile includes only the highest grades of coal
- Contract (large mines)	
- Spot (small mines)	. Utility Sector
- Surge (up to 25 million tons)	- Point estimates for KWH sales by region
. New Capacity	
- Based upon BOM demonstrated reserve base	- KWH sales allocated to four load categories (base, intermediate, seasonal peak, and daily peak)
- Reserves allocated to model mine types	- Existing generating capacity utilized by model on basis of variable cost
- Minimum acceptable selling prices estimated for each model mine type	- New generating capacity utilized by model on basis of full costs (including capital costs)
- Upper bounds of new mine capacity for each region based upon planned mine openings	- Air pollution standards addressed explicitly
. Coal washing	- Transmission links between regions
- Basic washing assumed for all bituminous coals	- Oil and gas prices fixed
- Deep cleaning option available to lower sulfur content to meet New Source Performance Standard or a one percent sulfur emission limitation for existing sources	- Coal prices determined from supply sector through transportation network
NON-UTILITY DEMAND	TRANSPORTATION
. Five non-utility sectors (metallurgical, export, industrial, residential/commercial, synthetics)	. Direct links
	. Cost based upon unit train or barge shipment rates
. Point estimates for Btu's demanded	
	. Lower bounds used to represent long-term contract commitments
. Allowable coals specified in terms of btu and sulfur content	
	. Upper bounds could be used to represent transportation bottlenecks or limited capacity
. No price sensitivity	

A summary of the spatial, temporal, and informational resolution of the CEUM is presented in Table 5. A listing of the model's important variables is given in Table 6. For further discussion and evaluation of the component parts of the CEUM see Reference (10).

Although not explicitly assessed in Reference (10), a particularly basic change was made in the CEUM starting with the first EPA study in the fall of 1977. Previously, each target-year's solution was derived independently of those for other target years. The model was revised so that runs for later target-years used earlier target-year results. Intertemporal constraints were incorporated in the following way: First, lower bounds were set on coal flows to insure that contracts undertaken would continue in force. Since it was assumed that 80 percent of sales were contract sales, transportation links and utility coal flows from coal piles to plant types within demand regions were lower bounded at 80 percent of deliveries in the prior target-year solution. Second, utility capacity additions in the CEUM consist of all plant capacity added since 1975. The modification of the model imposed lower bounds that required capacity additions by plant type in a later target year to at least equal those of the prior target year.

The next three sections of this paper focus on the linear programming formulation of the CEUM and issues relating to the overall structure of the model. By the use of an illustrative linear programming matrix, it will be shown, in general terms, how the CEUM's four major components interrelate.

Table 5

RESOLUTION OF THE CEUM

Spatial

 o 30 Coal Supply Regions

 o 35 Utility Demand Regions

Temporal

 o Static

  A Single 5 to 30 Year Time Block

Informational

 o 40 Coal Types

  5 BTU Levels and 8 Sulfur Levels

 o 5 Non-Utility Coal Consuming Sectors

Table 6

CEUM VARIABLES

Endogenous Variables

- o   Coal Supply
- o   Coal Cleaning and Mixing
- o   Coal Transport
- o   Oil/Gas Procurement
- o   Coal Procurement by Non-Utilities
- o   Electricity Generation from Coal
- o   Electricity Generation from Non-Coal Sources
- o   Electricity Transmission
- o   Building Electrical Generating Capacity
- o   Building Scrubber Capacity

Exogenous Variables

- o   Electricity Demand
- o   Non-Utility Coal Demand
- o   Bounds on New Coal-Fired Capacity
- o   Fixed Nuclear and Hydro Capacity Additions
- o   Bounds on Scrubber Capacity
- o   Oil/Gas Prices
- o   Capital Costs, O&M Costs, Transportation Costs, Etc.

V. Discussion of the Linear Programming Matrix

The linear programming (LP) matrix presented in Chapter 2 of Reference (10) illustrates the basic structure of and the naming conventions used in the ICF Coal and Electric Utilities Model (CEUM), for one supply region, Virginia (VA), and one demand region, Western Pennsylvania (WP). Each column in the LP matrix represents either a physical or notional economic activity. Positive entries in a column represent an input into the associated activity; negative entries represent an output of the activity. The last entry in each column represents the annualized cost of operating each activity at unit level and forms the coefficient of that activity in the objective function.

Nine major types of activities appear in the LP matrix. These are:

o   coal mining

o   coal cleaning

o   coal transportation

o   oil/gas procurement

o   coal procurement by non-utilities

o   electricity generation from coal

o   electricity generation from non-coal sources

o   electricity transmission, delivery, and load management

o   building electrical generating and scrubber capacity.

Each row of the LP matrix, except for the last row, represents a constraint associated with a physical stock (coal, heat energy, electricity, etc.) or, in some cases, with a consumption requirement. Physical stocks may be of fixed size, exogenously specified, or of variable size, created by activities within the model. Constraints associated with stocks of variable size are called material balances; they force quantities created within the model to equal or exceed quantities used.

Seven major constraint categories appear in the LP matrix. These are:

o   available coal reserves by mine type at supply regions

o   coal stocks by coal type at supply regions (material balances)

o   fuel "piles" at demand regions (material balances)

o   non-utility energy requirements at demand regions

o   electricity constraints, including electricity consumption requirements, and electricity supplies (material balances), at demand regions

o   electrical generating and scrubber capacity constraints, including fixed generating capacity for existing plants, material balances for capacities not yet built (new plants), and material balances for scrubber capacity on both existing and new plants

127

o   new capacity building limitations for generating electricity.

The following conventions have been adopted with respect to constraint rows in the LP matrix:

o   constraints imposed by exogenous size limitations of existing stocks are specified with positive entries on the right-hand-sides of the associated rows

o   material balance constraints are specified with zero entries on the right-hand-sides of the associated rows

o   constraints imposed by exogenous consumption requirements are specified with negative entries on the right-hand-sides of the associated rows

o   negative entries in a constraint row indicate additions to a stock; positive entries indicate subtractions or use.

The last row of the LP matrix designates the objective function. Its entries are the costs (1985 costs in 1978 dollars) of operating the associated activities at unit level. While the interpretation of most of these entries is self evident, we note that the objective function coefficients for the electricity generation activities represent annualized O & M costs for all plants (existing and new) except for nuclear capacity which is modeled with its annualized fuel costs as part of its O & M expenses. The objective function coefficients for all building activities represent annualized capital costs, where a real annual fixed charge rate of 10% is used.

In general, the various activities in the LP matrix have the following effects:

o   Coal mining activities transfer coal from available coal reserves to coal stocks at supply regions.

o   Coal cleaning activities transfer coal from a stock of one coal type to a stock of another coal type (possibly of lower sulfur level), allowing for cleaning losses.

o   Coal transportation activities transfer coal from coal stocks at supply regions to fuel piles at demand regions.

o   Oil/gas procurement activities place oil and gas in fuel piles at demand regions.

o   Coal procurement activities by non-utilities remove coal from fuel piles in order to satisfy exogenous non-utility energy demands.

o   Activities for electricity generation from coal remove coal from fuel piles, use electrical generating capacity and possibly scrubber capacity, and create electricity supplies.

o   Activities for electricity generation from non-coal sources remove non-coal fuels from fuel piles, use electrical generating capacity, and create electricity supplies.

128

o   Electricity transmission activities reduce electricity supplies in one region and increase them in another region, allowing for transmission losses. Electricity delivery activities reduce electricity supplies in order to satisfy exogenous electricity consumption requirements, allowing for distribution losses.

o   Activities for building electrical generating or scrubbing capacity create new capacities. Exogenously specified limits may be imposed.

VI. Discussion of the Objective Function

For illustrative purposes, a verbal representation of the CEUM's objective function is presented below. A more detailed mathematical representation can be found in Chapter 2 of Reference (10).

The objective function is an annual dollar cost measure that is minimized. It includes nine different types of terms. The first term multiplies real annuity coal prices by annual amounts of coal supplied in each supply region, at each cost-of-extraction level, at each BTU content level, and at each sulfur content level, to achieve a total coal production cost. The second term represents total deep cleaning costs for each supply region, at each BTU content level. The third term multiplies coal transportation prices by the amounts of coal transported annually between each supply and demand region, for each heat and sulfur content level. The fourth term is the product of prices and quantities of oil and gas consumed in each demand region.

The remaining terms of the objective function collect costs from the electric utility sector. The first of these terms multiplies appropriate operation and maintenance costs by the annual amounts of electricity generated in each demand region, for each plant type, fuel type, and load mode. The next term multiplies transmission costs for new lines by the annual amounts of energy transmitted via new lines between pairs of demand regions. The seventh term in the overall objective function is the product of electricity delivery costs and the annual amounts of electricity delivered in each demand region. The eighth term multiplies annualized capital costs for new plants by the amounts of generating capacity built in each demand region, for each plant type. The final term in the objective function is the product of annualized capital costs for scrubbers and new scrubber capacities (for each of four different scrubber-types) in each demand region.

An exact understanding of the types and implications of different terms in the objective function is of course a necessary initial step in the assessment process. We repeat that complete mathematical details can be found in Chapter 2 of Reference (10).

VII. Structural Issues

In this section, we shall discuss issues relating to the overall structure of the CEUM, and present a preliminary evaluation of that structure. The CEUM is structured in a straightforward linear programming framework, described in Section V. The model is large in size because coal supply, transportation, and use by electrical generating plants are all represented in great detail. Given this high level of disaggregation, only an extremely simple structure (such as the LP framework employed) would be sufficiently tractable to yield solutions with a reasonable computational effort.

In any model of this sort, there is a trade-off between the level of disaggregation (complexity of data) on the one hand, and the complexity of the structure, on the other. Given computational limitations, the more disaggregated the data, the simpler the structure must be. The choice of the best compromise between these two forms of complexity is a difficult one for modelers to make, and it is heavily dependent on the purposes for which the model is to be used. The ICF choice, highly disaggregated data with a very simple model structure, is near one end of the spectrum of possibilities. An evaluation of this difficult choice must be an important part of an assessment of the CEUM.

The simple LP structure and high level of disaggregation of the CEUM have a number of advantages:

o   The structure permits a "natural" representation of the energy sector of the economy. Almost every column of the LP matrix, as explained in Section V, represents a tangible economic activity. Once the notation is mastered, and the derivation of the data is understood, it is an easy matter to interpret any part of the model as a description of an economic process or processes.

o   New data, or new economic processes, are easily assimilated into this framework, so that the model can be readily modified or updated.

o   The ability to operate at a high level of disaggregation allows the representation of considerable regional detail, so that solutions of the model may have policy implications for specific regions.

o   Being highly disaggregated, the model is more stable and less subject to extreme corner solutions than smaller, more aggregated, LP models would be.

The simple LP structure of the CEUM also has some significant disadvantages:

o   As in all LP models, every economic process must be represented in exactly the same way: as a perfectly divisible activity that uses all inputs and produces all outputs in fixed proportions.

o   Any solution of the model must be the solution of a linear optimization problem, in this case the minimization of the total cost of specified electricity production and coal consumed in non-utility sectors. Although cost minimization characterizes   a   purely   competitive

131

equilibrium, it is far from clear that cost minimization is a characteristic of a regulated monopolistic industry such as electricity generation. In fact, it is doubtful that the behavior of this industry can be described by the solution of any optimization problem. None of the voluminous economic literature on the behavior of regulated utilities was or could be brought to bear given the LP model structure.

o   The model is completely static. All events must be collapsed into a single time period. Behavior that changes over time cannot be represented or described in the context of the CEUM. In a short-run analysis, for those aspects of coal supply, coal transportation, and electricity generation which can change but slowly, this may not be a serious problem. However, when the horizon of the model recedes to the year 2010 or later, the ability of the model to produce any useful results becomes suspect. A time period of thirty years or greater is sufficient for coal mines to open and close, for new technologies to come into play, for patterns of electricity use to vary, and for market conditions for alternative fuels to change, so that it becomes impossible to represent the distant future in a timeless model. In addition, it is unreasonable to represent the distant future in a deterministic framework. We believe that for such modeling, a highly aggregated, dynamic stochastic model is more appropriate.

o   Because the structure of the model is rather rigid, available data must be adjusted to fit the model. A model constructed to take best advantage of available data would have had to be more complicated and less structurally uniform. The CEUM required much data that was not available, so that the data had to be manufactured. As a result, much of the apparent detail of the model solutions depends on assumptions with little or no empirical basis.

o   The CEUM combines a very high level of detail in coal supply and electricity generation with a very highly aggregated, static description of alternative fuels including oil, gas, and nuclear fuel. This combination of aggregate and disaggregate analysis is especially dangerous given the simple analytical structure of the CEUM. Size and disaggregation of data are to some extent a functional substitute for complexity of structure. In a large, disaggregated model, the set of feasible solutions can be bounded to include only those with realistic and reasonable properties. In particular, the use of a large number of activities and constraints allows a linear model to approximate the behavior of a nonlinear one. On the other hand, if unrealistic results are to be avoided, it is often essential (and usually inexpensive) to give a highly aggregated model an explicitly nonlinear structure. Because the CEUM is intended to be primarily a model of the coal sector of the U.S. economy, it is not surprising that the coal sector is described in much greater detail than are alternative energy sectors. Unfortunately, the ICF modelers found it necessary to integrate both the coal sector and the aggregated alternative-energy sectors into the same rigid LP structure. As a result, there is little possibility of generating a realistic description of the alternative energy sectors within the CEUM. Because the coal sector is so strongly dependent on alternative energy forms, systematic errors are undoubtedly introduced into model solutions.

A summary of important structural issues is presented in Table 7.

Table 7

CEUM STRUCTURAL ISSUES

Advantages

- o   Simplicity of Structure
- o   Physically Significant Variables
- o   Easily Updated, Given New Data
- o   Regional Detail
- o   Stability - Due to Disaggregation
- o   Stability - Due to Data-Driven Accounting Structure

Disadvantages

- o   Complexity of Data
- o   Simplistic Nature (Linearity, National Optimum)
- o   Optimization Instead of Simulation (Regulated Industry)
- o   Static Nature
- o   Fitting Data to Model (Rigid LP Structure)
- o   Uneven Detail Across Energy Sectors
- o   Difficulty of Further Spatial Disaggregation

      (Data and Computational Limitations)

Having briefly analyzed advantages and disadvantages of the disaggregated LP framework used in the CEUM, what conclusions can we draw as to its appropriateness for the problems at hand? We believe that in the development phase of the model, the simplicity of the LP framework, and the ease of interpreting, modifying and updating it, more than compensates for its limitations. However, now that the model is reasonably complete and is being used for policy-making purposes, consideration should be given to embedding a highly aggregated variant of the present CEUM into a dynamic system. This dynamic version of the model could be run side-by-side with the more disaggregated static version to serve as a check on serious systematic errors in the latter. We believe that in model runs with long horizons (30 years or more) a dynamic model with more behavioral representations may be indispensable as a tool for generating constraints to be used by the static CEUM.

VIII. General Design of the CEUM and Strategies for Further Assessment

In this section we examine ICF's choice of a general design for the CEUM and discuss strategies for further assessment. The general design of a policy forecasting model strongly influences the properties that the model will have. Among the most important of those properties are:

(a) accuracy

(b) detail

(c) range of application, and

(d) generality.

The properties of 'accuracy' and 'detail' are self-explanatory. The 'range of application' refers to the number of issues which a policy model can address given a particular state of the world. The 'generality' of a model refers to the number of different states of the world for which the model is relevant.

In and of itself, high levels in each of these areas are desirable. However, given a fixed amount of resources for the development and operation of a model, compromises must be made so that a reasonable balance of the various properties can be achieved. For example, a high level of accuracy may require a low level of detail, while a high level of detail may preclude a broad range of application. Obviously, there is also a trade-off between range of application and generality.

The general design of a policy model should depend in large part on the desired mix of the above properties. From the nature of the CEUM it seems clear that ICF placed a high priority on achieving a high level of detail in their model. The range of application was intended to be broad but significant amounts of accuracy and generality may have been sacrificed.

Emphasis on detail was a natural and necessary choice. Coal is a very heterogeneous commodity in two different senses. First, there are many varieties of coal, each with different properties and uses. Secondly, coal is located in different places and transportation costs are high compared with the cost of mining and utilizing coal. A model which aggregated many types of coal into one classification, or which, through omission, failed to distinguish between different locations of coal deposits, would have a limited range of application, indeed. Clearly, to be broadly useful, a coal model must be reasonably detailed.

At this point two questions arise.

(1) Was the target level of detail for the CEUM the appropriate one given the overall goals of the model builders and the required sacrifices in other desirable model properties?

(2) Is that target level of detail actually attainable in the framework of the overall design of the model?

We are not far enough along in our assessment (nor would it be fair) to unilaterally comment on these points. Instead some indication will be given of the directions in which we intend to proceed.

A number of sensitivity analyses have been performed using the version of the model that we are assessing. The results of some of these studies are displayed in Figures 1 and 2. When these plots are digested they offer important insights into mechanisms that seem to be at work within the model. Such insights, together with ideas that have been gathered in studying the documentation and the computer code, have helped to formulate a set of issues that will be pursued. An initial plan illustrating how some of these issues might be studied is shown in Table 8. There are three types of model runs that will be made:

(1)   Equivalence Runs - to establish that the version of the model that is in-hand corresponds to published numbers,

(2)   Screening Runs - where groups of issues are simultaneously explored to quickly gain insight into interesting or counterintuitive results, and

(3)   Issue Runs - to examine critical aspects of specific issues.

The development of a strategy concerning the number, types, and order of the model runs that will be conducted represents the current stage of our assessment of the CEUM.

Figure 1

MAPPING OF SENSITIVITY STUDIES

Utility Oil/Gas Consumption

vs

Coal-Fired Generating Capacity
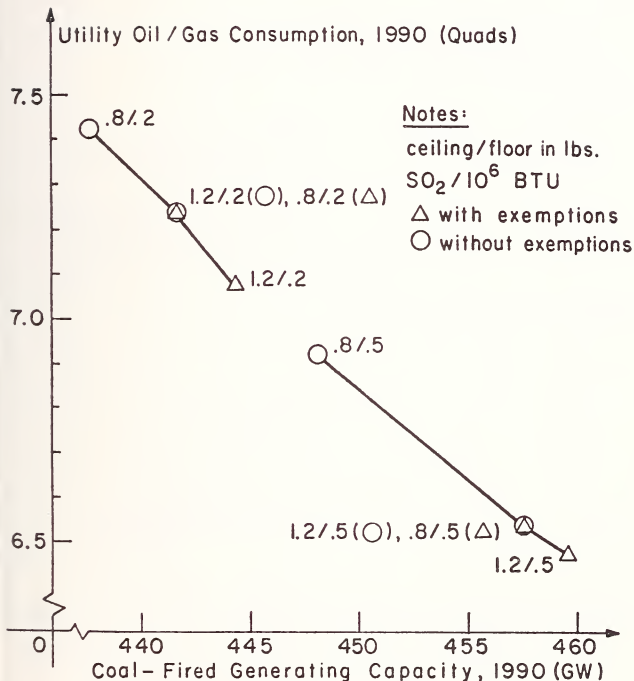
Figure 2

MAPPING OF SENSITIVITY STUDIES

Western Coal Produced for Eastern Consumption

vs

Midwestern Coal Production
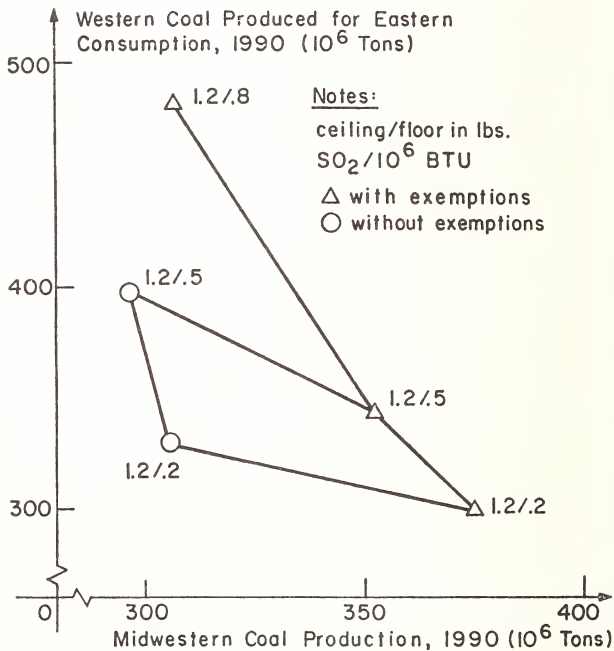
Table 8

PRELIMINARY STRATEGIES FOR SENSITIVITY ANALYSES

1. Base Cases
   - o   Scrubbers Mandatory:
         Yes (ANSPS), No (NSPS)
   - o   Floors (lbs. sulfur dioxide per million BTU):
         .20, .50, .67, .80, 1.2
   - o   Ceilings (lbs. sulfur dioxide per million BTU):
         .80, 1.2
   - o   Exemptions:
         With, Without

2. Increase Scrubber Cost:
      5%, 10%, 25%

3. Upper Bound Scrubber Capacity:
      10^6 GW (Base Cases, Unbounded), 100 GW

4. Lower Bound Appalachia Coal Production:
      400 Million Tons (1990) with Scrubbers not Mandatory and with
      Stricter Sulfur Standard than NSPS

5. Mine Lifetime (Years):
      30 (Base Cases), 20, 40

6. Electricity Demand Growth Rate (1985-1995):
      4.0% (Base Cases), 5.0%, 3.0%

7. Nuclear and Hydro Capacity Additions:
      25% Decrease from a Base Case

8. Non-Utility Coal Demand:
      10% Increase over a Base Case

9. Oil/Gas Prices:
      25% Increase over a Base Case

IX. Acknowledgments

X. References

1. Coal Supply Analysis, Prepared for the Federal Energy Administration by ICF, Inc., May 1976.

2. Review of Federal Energy Administration National Energy Outlook, 1976, Prepared for the National Science Foundation by Resources for the Future, March 1977.

3. Economic Analysis of Coal Supply: An Assessment of Existing Studies, Prepared for the Electric Power Research Institute by Pennsylvania State University, Principal Investigator: Richard L. Gordon, EPRI EA-496, Project 335-2, July 1977.

4. The National Coal Model: Description and Documentation, Prepared for the Federal Energy Administration by ICF, Inc., August 1976.

5. Coal and Electric Utilities Model Documentation, ICF, Inc., July 1977.

6. Coal in Transition: 1980-2000, Energy Modeling Forum, EMF Report 2, Stanford University, September 1978.

7. Effects of Alternative New Source Performance Standards for Coal- Fired Electric Utility Boilers on the Coal Markets and on Utility Capacity Expansion Plans, Prepared for the Environmental Protection Agency by ICF, Inc., Draft Report, September 1978.

8. The Demand of Western Coal and its Sensitivity to Key Uncertainties, Prepared for the Department of Interior and the Department of Energy by ICF, Inc., Draft Report, June 1978.

9. Further Analysis of Alternative New Source Performance Standards for New Coal-Fired Power Plants, Prepared for the Environmental Protection Agency and the Department of Energy by ICF, Inc., Preliminary Draft Report, September 1978.

10. The ICF Coal and Electric Utilities Model: An Overview Assessment, Prepared for the Electric Power Research Institute by the MIT Model Assessment Group, MIT Energy Laboratory, Energy Model Analysis Program, February 1979.

DEVELOPING, IMPROVING AND ASSESSING
THE ICF COAL AND ELECTRIC UTILITIES MODEL

C. Hoff Stauffer, Jr.

ICF Incorporated
Washington, DC

When I sat down to sketch out some notes for this talk, I wrote at the top
of the paper, "Model Validation and Assessment." I didn't notice until today
that my talk was supposed to deal with the ICF Coal and Electric Utilities
model.

Fortunately, one of my conclusions was that the models ought to be vali-
dated and assessed using the same process that we actually used to develop the
ICF model, and that we use continually to validate it and assess it for our
own purposes. In other words, I believe the model development and model
assessment should employ the same analytic steps.

I should make it clear that I view myself as a model user, not a model
developer or a modeler. Similarly, I have a limited experience with models.
The only models with which I have dealt have been structural, not economet-
ric. I think many of my comments would not apply to econometric models.

Definition of Assessment and Validation

When I began to organize my thoughts on this topic, I realized I didn't
really know what was meant by "assessment and validation." After listening to
the talks so far today, I still don't know what others mean when they use
these and related terms. However, I developed an operating definition for
myself so that at least I know what I mean.

Put most simply, I decided "assessment and validation" must have to do
with whether the model gets the right answer.

Then I note there are two parts to that. One part is whether the model
measures relative changes correctly. The federal government is generally con-
cerned with relative changes. The second part is whether the model measures
absolute levels correctly. The private sector is generally most concerned
with absolute levels.

Finally, I note there is another important dimension as well. One must determine for what <u>kinds of questions</u> the model gets the right answer, and for what kinds of questions it does not get the right answer.

Use of Historical Data for Validation

Now, if our definition of model validation is whether or not it gets the right answer, how can we <u>determine</u> whether it is capable of this? First of all, we cannot use historical data. Most structural models involve investment decisions, and these involve lead times and expectations. The expectations being modeled were never written down.

As an example, if we want to predict a decision someone would have made in 1965 regarding a powerplant which would come on in 1975, the last thing I would want to use would be actual historical oil prices. I would want to use the oil prices which the decision-maker thought were going to exist. But there is no way to know that precisely.

So we discover that "backcasting" is no easier than forecasting. In backcasting and forecasting, you need to assume expectations. There is no comprehensive data source for expectations. So I think it is clear that backcasting is not a useful approach to model validation.

Model Comparisons

Another way to determine whether a model is capable of getting the "right" answer is to compare its answers to those of other models. But this seems to me to have limited utility. If we wanted to decide things by consensus, we could ask people to vote. More importantly, if the answers <u>are</u> different, what does that prove?

Unless model comparisons proceed to the seven steps I outline below, I think it is clear that such superficial comparison exercises--where only outputs are compared--are effete endeavors.

Develop Confidence In Model

These two deadends--backcasting and model comparisons--lead me to a third approach. This approach is not simple nor quick, nor is it mindless as the first two approaches. This approach requires intellectual capital, time, and hard work. This approach is to do the analysis required to develop confidence in the model.

This in turn requires an in-depth understanding first of the phenomenon being modeled, the issues the model is designed to address (or the question it is designed to answer), and the dynamics of how the issue areas affect or are affected by the phenomenon being modeled. It also requires an in-depth understanding of the structure of the model, its data, and its assumptions. Finally, it requires experience with the model and careful analysis of its forecasts.

142

The process we used to develop our model and refine it, and the process which, in my opinion, people should use to assess and validate other models, as seven steps.

Step 1: Understand Phenomenon to Be Modeled

The first step would start off by understanding, in great detail, the phenomenon that the model is trying to deal with. In the case of the ICF Coal and Electric Utilities model, these phenomena are the coal and electric utility industries.

The richness of detail necessary for a complete understanding of the subject in this case begins with a knowledge of coal reserves: the quantity, quality, and physical characteristics thereof, all by geographic region. And, by the way, each of those dimensions has several sub-dimensions.

Similarly, you have to understand mining; you have to understand technology and cost and environmental regulations. You have to understand the economics of mining and how a producer views an investment to open a mine. You have to, and I'm going to come back to this, understand what is meant by price. There are many definitions of what seems to be a simple number price. Only the definition is correct for any one use of it.

Then you have to know about coal preparation, the effect it has on coal quality, the cost thereof, and the tradeoffs. You have to understand that coal transportation, its various modes and their costs vary by geographic region.

You have to understand differences between consuming sectors: utility, industrial, metallurgical, export, and so forth. Within a sector, you have to understand the combustion trade-offs, particularly in the utility sector. You have to understand how electric utilities dispatch their capacity to the daily load curve. You have to understand how they plan capacity expansion.

You have, therefore to, understand power plant costs: both capital and operating. Very importantly, you have to understand environmental regulations, and the cost of complying with them. You have to understand transmission. You have to understand the difference between the long-term dispatch kind and the short-term dispatch kind of models, and the effects each would have on the kind of decisions you'll get.

You have to understand finance very well. Financial considerations have overwhelming influences.

You have to understand the nature of the coal markets. By that I mean, in this case, there is a long term contract market and there is a spot market; there is very little in between. The market varies by geographic region. It varies by type of coal and by sector. Even the kind of firms that produce coal for those different markets are different.

143

In summary, you must understand in great detail the substance of what is being modeled if you are to be capable of judging whether the model is properly structured, has valid data, is based on reasonable assumptions, and finally gets the right answer.

Step 2: Specify Issues to Be Addressed

The second step would be to understand the <u>issues</u> which the model was designed to address, or the questions the model was intended to answer.

No model is going to be designed to answer every possible question, so this must be specified beforehand. The ICF Coal and Electric Utilities Model (CEUM), for example, is a long-term model with a high level of detail, geographic and otherwise, in the coal and electric utility sector. This means that the model can address such issues as the effect of alternative new source performance standards on the coal and electric utilities industries. It can forecast emissions, it can forecast costs, it can forecast oil consumption, and it can forecast regional coal production. It can do leasing analyses, and it can do some tax and rebate analyses. As it turned out, we can use it to evaluate new technologies such as coal slurry pipelines. It can do basic sensitivity analyses like changes in severance taxes, wage rates, productivity and other coal parameters.

It will not handle short-term issues, or transitional issues such as the expected shutdown of considerable mining capacity in the near future as a result of the soon-to-be-promulgated strip-mining regulations. It won't deal with strikes. It won't deal with bad weather. It won't deal with rail car shortages. This is because it's a long-term model.

Similarly, it won't deal with broader issues. It won't deal with, for example, the trade-off between electricity and gas for home heating, and it won't deal with the effect of reduced U.S. oil imports and world oil prices.

Some of these concerns may be self-evident; but the point is that the model is designed to answer some questions and not others. Before we even start to look at the model structure, we must understand what it is trying to do, what it was designed to do.

Step 3: Understand Dynamics of System

Then the third step, still before we examine the model itself, is to understand how the issues the model is designed to analyze might affect the phenomenon being modeled, in this case the coal and electric utility markets.

As an example, if EPA were to make more stringent the new source performance standards for coal-fired power plants, we could expect that emissions on the new plants would be reduced and that the cost of the new plants would be increased. We would expect to see a shift to higher-sulfur coals. We would expect to see a change in the dispatch order. Most models won't allow the dispatch to change in response to the economics. In this particular new source performance standard problem, that is critical.

Stricter new performance standards could, for example, help create a shift to oil; as the costs of operating a coal-fired plant increase with more stringent emission standards, oil plants would become comparatively more economical. We could expect fewer shipments of Western coal to the East. Regional emissions could be expected to decrease generally, although the changes in emissions and cost by region may be very different. Power plant reliability could be expected to change, and that would affect the economics.

These are the kinds of changes one might expect to result fom a change in the new source performance standard. If the model was designed to address the new source performance standard issues, it should be structured to measure such changes.

Step 4: Analyze Structure of Model

Once we have examined these preliminary considerations, the fourth step then is to look at the model itself and ask whether it is structured to pick up the kinds of dynamics we've been discussing.

For example, one of the first direct impacts we would see under the stricter standards discussed above is that a new plant is going to cost more. Its emissions rate is going to be reduced. Can we enter that in the model as a clean number in the form of an emission rate or a cost? Or must the model be modified in some way, such as the price differential between highand low-sulfur coal up to 40¢ per million Btu? When we do something like that, we're assuming an answer, not doing an analysis.

So, in evaluating the model, the first question is whether these direct effects can be entered as clean numbers, or whether an answer must be assumed.

The second question is whether the model is structured to pick up the dynamics and the trade-offs that we know are going to occur.

One of the dynamics mentioned earlier, economic dispatch, is critical for most coal and electric utility analyses. The model has to be able to dispatch generating capacity by region on an economic basis. That is how it is done in the real world.

Similarly, because we know the policy people are interested in Western coal production and how much coal comes East, we must know whether the model is set up to measure that with any precision. That turns out to fall heavily on regional disaggregation. If we represent the Midwest as a single geographic point, then the model will show that all a category of plants in the Midwest either will go to low-sulfur coal, or will not. In actuality, we know from other analyses that that line is going to shift back and forth between Iowa and Ohio, depending on circumstances. Thus, maybe Illinois will go to Western coal, but Indiana won't. Maybe western Kentucky will go to Western coal, and eastern Kentucky won't.

We need to know whether the model is set up to catch such specifics or whether it must represent that whole area as a single point, say, Indianapolis.

Finally, we need to know whether the model is structured to provide the output that we want. For example, we can measure emissions and costs by region, or regional production by sulfur contents? Are the outputs there or only some hints from which, as Peter House says, the forecast is created on the typewriter? We need to know these specifics.

This, then, is the fourth step: given an understanding of the industries, given an understanding of the questions we are trying to answer, and given an understanding of how these issues affect the industries, then, we need to know whether the model is structured appropriately, on the input side, on the internal structure side, and on the output side.

I think this is the key point in most assessments. If it can't pass this test, one might as well not proceed. Also, if one can't tell from the documentation, then the model similarly gets low marks.

Step 5: Review Critically Data Inputs

The next step after the structure is the data inputs. The first thing we would look for is whether the data are well-documented. By well-documented I don't mean saying that they got the power plant capacity from the FPC. Anybody who understands FPC power plant capacity data knows there is a great deal of noise there.

What the documentation should say is that the power plant capacity came from the FPC and probably 30 other sources. I think that the actual numbers used should be there, because there is so much room for interpretation. There is so much noise in this kind of data, that the user has to be able to tell whether the mistakes were caught in the FPC data or whether EIA data mistakes that we know of are there.

And incidentally, I want to distinguish between models and modelers or analysts. There is a difference between them. A model by itself is fairly irrelevant. Part of evaluating the model is making sure that the people involved have taken adequate care of the data. There should be clear statements of the strengths and weaknesses. The modelers ought to furnish their judgments concerning whether the data they used were any good. They should state what data they think are critical and what difference it makes. I worry about people who call themselves modelers, creating beautiful structures, and ignoring the data. Because a good structure will depend on the data.

Step 6: Critically Review Assumptions

After we understand what we are trying to do, first we look at the structure, then the data, and finally the assumptions.

146

The assumptions should be clearly spelled out. They are usually called scenario specifications these days. Whatever they are called, they should be clearly documented. They should be accompanied by a discussion of which ones make a difference, and what kind of difference they make, and which ones are most critical, and which ones are least certain.

None of the assumptions should be omitted, and they should all be presented right up front. They should not be buried in the documentation.

Step 7: Analyze Model Forecasts

Thus far in the process of evaluation, we have gone through six steps without even looking at the model itself, without looking at any of the computer code or any of the forecast measures. We are now ready for the **final** step, which is to examine the model. This is not the first step; it is the last step.

As we begin this step, we should approach the model from the context of an issue. We don't want to examine the model in strictly abstract terms. Rather we acknowledge that it was designed to evaluate a given set of questions or issues, and then we analyze its behavior on that issue.

In the third step we defined what we expected to happen when certain policy alternatives were played out. Now we want to see whether the forecasts behaved as we expected. If not, we should be able to explain why not. For example, one analysis of the new source performance standards involved forecasting what should happen if EPA set up the most stringent standards it could devise. In general, we would have expected this to result in lower emissions. But the opposite turned out to be the case. The explanation was that if the cost of a new plant increased, economic dispatch would cause the existing oil and coal-fired plants to take on higher loads. These plants have very much higher emission rates than the new coal plants. Therefore, the system would end up having higher emissions than otherwise, because the increased emissions from existing plants more than offset the reduced emissions from new plants.

This was an important finding and one which we had not expected. The richness of the input data and output data from our model provided us the means to understand this anomaly. The sound structure of the model resulted in an effect that was not anticipated but now is believed to be valid.

One other point, which relates again to the documentation, is that one should be able to reproduce any forecast variable, using documentation, other forecast variables, and an understanding of the model structure. I consider this a critical point. This is an important way we can build confidence in a model. It lets us know whether it's actually working the way it should.

If one can do this--and I and many of our clients and I can do so with ou
model--there is no need to review the computer code. One can determine
whether the code is correct by analyzing the inputs and outputs. For those
who are interested in the code--it can be reviewed. But such review is not
necessary. I've never looked at the code to our model, and I'm convinced I
understand thoroughly how the model operates.

Now, this seven-step process which I have defined is the process we used
to build the model, and the one we apply for evaluation as we continue to
refine the model.

Expected Findings of an Assessment

I want to summarize what I would expect to be the results of an assessmen

The first thing such an assessment should deal with is the quality of the
documentation. This is the basis upon which people are asked to believe the
answers. In order for them to be able to go through the process we've just
outlined, they must see the documentation as their vehicle.

I should distinguish here between two types of documentation: one is
aimed at the user, and the other at the operator. The first is keyed toward
the person who will use the results, possibly to make decisions. The second
is keyed to the operator, who wants to take the model and make it go. I
believe the first type is overwhelmingly more important than the second.
Indeed, for my own purposes--since I never intend to operate the model--the
second is irrelevant. At any rate, the two types are distinct and separate.

Incidentally, there are examples of models used in government whose docu-
mentation is extremely inadequate. If we look at such a model and, as part o
an assessment, attempt to duplicate a certain number from the model, we might
find it impossible. When this happens, it is a failure of the documentation.
And it happens frequently. The excuse given is that the modelers are too
busy running the model to go back and document it. But the model just isn't
very useful unless an outside assessor can determine whether he can believe i

Secondly, an assessment should show what kinds of questions the model is
designed to answer, and the kinds it is not designed to answer, and maybe som
that are in between--that is, perhaps some that the model could answer if cer-
tain improvement opportunities were taken.

The assessment should deal then with the quality of the forecasts. What
are the strengths and weaknesses of the forecast we get for the questions
which the model is designed to answer? Will it measure relative changes? Ho
well? The structure is the key element in measuring relative changes. The
data are the key elements for the absolute levels.

I think it is wrong to review a model in the abstract. It should be reviewed in the context of what it was designed to do, or one of the things it was designed to do. I think looking at the real-world problem helps to focus one's thinking on what is important and what is not important.

If the assessor wants to suggest improvement opportunities, I think they should be done within some perspective. They might say what could be done, indicate the benefits of doing it, assess the feasibility of doing it, and then estimate the cost of doing it. People have suggested, for example, that our model ought to incorporate ash. I know that would be very expensive, however, and I can see very small benefits. The only appropriate way to suggest this kind of improvement is after having analyzed the benefits and the costs.

Also, I don't think it is reasonable to review a model alone. I think that is a silly thing to do. A model ought to be reviewed along with the analyst or group of analysts. Let me use the example of my Texas Instruments calculator. It is a very fine piece of technology. Give it to me, and with a little bit of time, I can compute a present value. My four-year old son is very bright, but give it to him, and he could not. It would be inappropriate to evaluate the calculator based on my son's performance.

I think that kind of analogy holds with models. Models are tools. In the hands of some craftsmen or analysts, they can be useful. In the hands of others, they may be useless or even destructive, if they become misleading.

Common Mistakes

Although I don't have time to dwell on them, I want to mention two common mistakes in modeling. One is the definition of the term "price." I think that people do not define price well; they often mix time periods and other options; thus, the assessor should be very careful about the definition price.

The same is true for "inflation." The problem is not as serious as it used to be. But, very often inflation, and its effects on the financial variables, are not treated consistently within a model.

I will end with something on a light note. Our model was used, as I mentioned, on the new source performance standards study for DOE and EPA, whose purposes and interests often conflict in that arena. Soon the environmentalists began to use it, and the industry began to trust it. I was very pleased that our model and our use of it was accepted by all these groups. That was a goal of mine, to create a tool and a reputation, so that those divergent groups would all be willing to use the same calculator, as in the analogy I used earlier. And that happened.

However, I think things got carried away, because we went through three phases of the analysis, and in the last phase we had 22 scenarios. I thought it was best summed up by a reporter, who called me after the last time these numbers were presented. He was asking some cogent questions, and that had seldom happened in the past, so the process was working. But then at the very end, he asked a very good question. He said, "Are you going to make any more runs?" I said, "No, I don't think so." He said, "Thank God!"

DISCUSSION

Mr. Ford (Los Alamos): I have a question for Hoff. And it is in regard to the remark that, in attempting to reproduce historical behavior, you need to know the expectations--in this case the expectations of the electric utility officials--and since that is not known, perhaps, the exercise of reproducing historical behavior could be skipped. And that line of reasoning could apply to almost all models and, therefore, one might say, well, we will skip this particular test, and all attempts to increase our confidence in a model.

I would suggest, if you can't get good information on what people antici-pated for the price of oil, and so forth, you demonstrate a set of expec-tations that, when fed into the model, create the historical investment decision, and then show those expectations so people can look at them and say they seem reasonable.

So, for a model that said electric utility officials expected the price of oil to go up by ten-fold in the next five years, one might suspect that that was an input that was jimmied to get the right investment decision. So, that would provide me, if I were looking at the model, one more test to look at to see how much confidence I could have in the device.

Dr. Stauffer: I think that is a good idea. But my comment was that you can't use actual historical data to feed the model. You must use expectations. You don't know expectations, so, you have to estimate them. Therefore, estimating the past is not necessarily any easier than esti-mating the future. But you have an interesting idea. I think that would be fun. It is a good point.

Dr. Nissen (Chase Manhattan Bank): Hoff, I would like to ask a leading question. And I am sorry that Lincoln Moses has just left, because he was the intended audience, but perhaps, the record can show the question.

One of the things that you have done is to provide us with a very impressive list of the kinds of data that have to go into even a piece of an energy system's model at the kind of level that generically we are talking about.

Not simply data about reserves, and so forth, the kinds of data that the constituency of the Bureau of Mines is used to responding to. But, data about costs, measurements of price, economic quality impacts of benefici-ation and preparation, data about transportation costs. And then when you get into utilities, you really get into the hard data--data that is, what you might call, high analysis content data. It is really not data that is recorded by a form, but it is data which is the output of an analysis process itself. Operating costs, environmental regulation impacts generating transmission distribution, scrubbing performance standards and impacts, and so forth and so on. The question I have is how much help was the data side of EIA in producing the data which went into your model?

I ask this, remembering the fact that we were all very proud in 1974 that Eric Zausner's group, at the time he was an assistant administrator, was called data and analysis and that was to bring about a wonderful symbiosis. I also remember how it looked, four years later, when I left.

The second part of this leading question is, how responsive do you anticipate the EIA data side will be in the effort to respond to deficiencies in the data, as it is collected in the near future? That is, is there any substantive interaction between you and the so-called data groups within EIA?

Dr. Stauffer: The last question first. How responsive do I think they will be? I just don't know. Within the last six months or a year, with one exception, there has been essentially no interaction between us and them, but there may be interaction between the analytic part of EIA and the data part, and that I don't know about.

On the question of how much, what value was all that data they collect. The answer is some. It has evolved over time. For example, for reserves, we used to use their reserve data exclusively. We are getting to the point now where we are going to the raw geology reports and modifying and adding to that data base. On the power plant, the analytic side -- things like capacity -- where we are on that is that we used to use their data, then we concocted what we call a master list, then we compare that master list to every other data source we ever see. When it is different, we call the power plant directly.

So, we think we now, and we call it our own, have the most updated variety of that. Lots of the inputs to the model, however, are not historical data or measurable things. They are engineering estimates. Like how much does a new power plant cost? And that kind of a number usually comes out of an analysis shop, or a technology shop, or they...

Dr. Nissen: You mean traditionally it has come out of an analysis shop.

Dr. Stauffer: Traditionally it has done that.

Dr. Nissen: What we can record is that the primary information side of the Energy Information Administration is providing almost no information to the analysis function.

Dr. Stauffer: Well, you said that.

Dr. Nissen: Excuse me. Institutional imperatives are to provide information to the cops, but not to the analysts.

<u>VALIDATION: A MODERN DAY SNIPE HUNT?</u>

<u>CONCEPTUAL DIFFICULTIES OF VALIDATING MODELS</u>

Peter W. House and Richard H. Ball
U.S. Department of Energy

Search for a Valid Model

Several years ago, one of the authors wrote a paper entitled, "Diogenes Revisited, the Search for a Valid Model."[1] Later, with John McLeod, this theme was taken and made a chapter in a book on large-scale modeling.[2] Rather than repeat those arguments given, let us capsulize and extend some of them here. In addition, we want to discuss a slightly different perspective which tries to suggest approaches to validation. The arguments can be focused on the following eight areas:

- Social science models often deal with phenomena at an empirical level where immutable natural laws cannot be ascertained; therefore, historical validation, even if possible, gives limited confidence that the resulting model has validity for predicting the long-term future.

- The formal statistical techniques used for validation are based on assumptions about the nature of the sample, such as the assumption that the future has some known relation to the past; hence, they may have difficulties similar to those of historical validation.

- There is little agreement as to what it means to be able to predict the future, with or without formal models.

- We are still very much in our infancy when it comes to measuring the state-of-the-present, using such techniques as indicators; consequently, we are hard put to say whether we have reasonable gauges with which to measure the future.

- Complex models are harder to validate than simple ones; but for most modern day problems, more complex approaches are necessary for policy analyses.

- Validation must be considered in relation to the type of model and to the purpose for which the model is used. Each combination has different implications for the feasibility, appropriateness and specific technique of validation.

- Models used to aid decisions and policy analysis should be judged on the basis of their utility in aiding decisions relative to alternative procedures, rather than on the same basis as models used in science.

- There are risks in insistence on validation, since inappropriate application of validation could unfairly discredit models that have real utility.

---

[1] House, Peter W., <u>Diogenes Revisited, the Search for a Valid Model</u>, March 1974, Washington Environmental Research Center, Office of Research and Development, U.S. Environmental Protection Agency, Washington, D.C.

[2] House, Peter W., and McLeod, John, <u>Large-Scale Models for Policy Evaluation</u>, New York: John Wiley and Sons, Inc., 1977, pp. 66-75.

The concept of validation, particularly validation of models which are used to predict the future, can be seen then to be based on very shaky theoretical grounds. Basic questions as to whether it can be done and why it is necessary to do it underlie the issues related to when we should try to validate models and how to do it.

Let us, for a moment, look at the question of when models can be validated from a different perspective. Imagine an n-dimensional array or vector. The axis for each dimension is so designed that it measures relevant variables describing models on scales from easy to hard, or small to large. Dimensions might represent size, type of model, subject matter, time horizon, and number of variables.

- Size would refer to the number and type of equations in the model and would suggest that the more equations a model has, the more complex it would be.

- Type of model--from pure simulation through the various optimization techniques. The closer the model tries to come to a solution to a problem, the more complex it is assumed to be.

- Discipline--in terms of the ability to measure and manipulate variables in each field. This is inherently very hard to portray on a single scale, in part because many policy models being built today utilize a mixture of disciplines. Perhaps a scale could be designed which would array disciplines according to their dependence on human behavior, since changes in human behavior and institutions present some of the greatest difficulties in predicting the future.

- Time horizon--the distance into the future the model is expected to forecast. Ignoring for the moment whether any forecasting exercise can be said to have validity, there usually is more information as to what the near future is apt to be like than the distant future, especially in technical terms.

- The greater the number of variables, the more complex the model is, the bigger it is, and the more difficult it is to assure the quality of the data.

There are several other dimensions that could be mentioned, but our discussion be carried out with these. Validating models is hypothesized to be a function where models are located in the space created by these vectors. As a general tendency, the closer the model lies to the origin, the more readily it will lend itself to standardized validation techniques. In other words, the simpler, more transparent the model, the easier it is to validate. This hypothesis does not in any way speak to the utility of the model, or to the efficacy of the validation exercise itself.

Validation itself has several facets and encompasses many techniques. We assume that validation generally pertains to a comparison of the model with reality. However, validation is variously interpreted to include partial validation of individual model algorithms, validation of input-output transformation properties and validation of overall output results. In the last of these interpretations, the extreme or "complete" validation position is that all parts of the model,

cluding input assumptions, should be considered part of the model and the
nal results or predictions should be compared against the real world.

 use validation to mean the general process of comparing model results to
ality, whether contingent on inputs or not. We believe that the appropriate
pe of validation depends both on the type of model itself and the application
r which it is used.

me models do not really require validation. If we use a computer model simply
 an accounting system to keep track of a number of variables, there may be
ttle to validate. Perhaps one can show that the control totals are consistent
th values obtained otherwise in a base case, but even that is not always mean-
ngful.

other example is where a computer is used to combine a large number of well-
tablished relationships in a simple way, i.e., where there is little chance
t the act of combining the relationships can alter their validity. A phys-
l example of this would be the modeling of a large number of objects that
ey Newton's law. Here one may argue that the algorithms have been validated.
 there are sources of error in the basic information or algorithms, then one
ght estimate the possible error in results as an alternative when direct
idation is not possible.

me types of model cannot be validated because they do not purport to predict
lity. An optimization-model in the normative sense might be validated in terms
 its causal algorithms, but it is rare that the optimality of the solution can
 verified independently. On the other hand, models that ape reality such as
mulation models, lend themselves most clearly to validation. This would include
timization models when the optimizing process is assumed to represent the real
avior of the system.

bject matter and discipline also make large differences. In the natural sciences,
ervations or experiments are charted and compared to a law or a set of known
ctions. The comparison of the observed to known relationships forms the basis
h for validating the experiment and for integrating the results with a body of
wledge. This latter approach provides a paradigm for at least some of the
idation work in other applications of modeling.

odel as an Experimental Device

lidation must be considered in relation to the use of a model. One of the uses
 a model is as an experimental device, in a series of "what if" exercises.

 first question in regard to validation is how we regard the role of the model.
 we wish to treat it as though it is a faithful replica of reality, then we should
idate it to confirm its practical utility. If we wish only to use it as a tool
 generate hypothetical scenarios, then perhaps no validation is required: the
usibility and self-consistency of the scenarios is what counts, rather than the
ns by which they are derived. If the model is itself the scientific hypothesis
 be tested, then validation is a primary objective.

In the case of a model assumed to represent reality for testing "what if" propositions as inputs, there is special concern if the model is based on empirical evidence where the underlying "natural" law may be unknown. Let us assume first that the model in question is one which concerns an institutional, organizational, or social structure. During the period being covered, there is likely to be a number of times in which the situation described will react to an event or series of events. The greater the number of variables being modeled, the more likely this will be. An economic model will have to take into consideration factors normally exogenous to the system being modeled, but which are turning points affecting all subsequent events. World War II and OPEC are examples. A model calibrated to truly ape these happenings would be one which was a simulation of a series of absolutely unique events. This situation differs dramatically from the natural law hypotheses which allow validation in the "hard" sciences. The use of such a unique event model to test "what if" hypotheses is open to serious question, particularly if there is no opportunity for the model to simulate behavioral reaction to situational change. Small perturbations in noncritical areas for very-near-term forecasts might have some credibility, but any more aggressive tests would be lacking.

A second use of model validation is to provide a basis for forecasting the future. Many of the problems alluded to earlier apply here also. A predictive model can be tested hypothetically to verify whether it performs as designed: it can be compared against historical data and then validated against its short-range predictions, if any, depending on the patience and objectives of the modeler. Although validation and historical comparison are not useless exercises, they are a far cry from claiming that the model is proven to have any prophetic capability. Knowing that the model operates as advertised means only that in the hands of competent analysts, the impacts of varying certain variables can be explained to other analysts in terms of what might have happened--had such changes occurred in a society much like the one we know, given extrapolated changes in all other variables. Helpful perhaps in decisionmaking, but a poor substitute for a crystal ball. Perhaps the best use for historical validation is to provide a baseline projection of the future. Then we can test how alternative assumptions about human behavior and events would modify the future.

Size of Model

Large-scale models have other problems which make their results difficult for policymakers and others to use with confidence. Several large-scale models are, for instance, really large in terms of numbers of variables. Recently, we estimated that the Strategic Environmental Assessment System (SEAS) has over 100,000 variables.3/ A model with this many variables causes obvious statistical problems in estimating the uncertainty. For example, if error bounds were assigned to each variable as though independent and their effect propagated through the system, the results of a given model might be estimated to be in the noise. In reality, actual

---

3/ A model used by both the Environmental Protection Agency and the Department of Energy for policy analysis.

156

ors might be greater if correlations exist or they might be less, especially
there are compensating effects. It would be desirable to take such correla-
ns into account, but in practice this usually is not feasible unless there
conservation principles or other constraints that limit the errors. It
ains to be shown that we understand the nature of individual uncertainties
ugh to deal with them properly. Of course, the important question is not
ther one can predict and validate the absolutes, but whether the predicted
nges due to policy initiatives are in the right direction.[4]

ardless of whether one agrees or not with the philosophical proposition that
idation is impossible on the basis of logic alone, sheer size would limit the
ability. Testing or applying existing statistical tools with a very large
el is a design problem which has yet to be solved.

k of Large Models

sibly the greatest hurdle to the validation of large-scale models is that, in
urist sense, there are almost none of them in existence. For example, in the
, of the three large models used for policy purposes (PIES, SEAS, and the
okhaven model), only SEAS can really be run as a total machine model.[5] The
er two require the intervention of analysts between the running of submodels
massage the data in order to get a full run of the system. The interjection
human analysts in the forecasts means that any form or mechanical validation
a snipe hunt--a technical impossibility. On the other hand, SEAS relies so
vily on exogenous data and assumptions that its forecasts can scarcely be
d to check on the model per se, while the input assumptions are too numerous
ever be realized simultaneously.

a final note on the number of variables, we should distinguish among input
iables, intermediate model variables, and output variables. When the number
uncontrolled input variables becomes large, validation becomes difficult in
nciple simply because the chance is small that all input assumptions will be
lized, i.e., the result depends more on the exogenous assumptions than on the
el.

number of intermediate variables in the model determines it complexity and
ects the ability to estimate errors, but it does not affect validation directly.
number of output variables determines the amount of data with which a model
ht be compared. A large number of variables makes the process more cumbersome,
it may render validation more feasible if the data is available. One might
n conjecture that validation is more feasible, albeit more work, when the ratio

---

Alonso, W., "Predicting Best with Imperfect Data," _American Institute of Planners_
Journal, July 1968, pp. 248-252.

See the earlier mentioned Chestnut study for a brief description of the models
or the more recent Greenberger, Martin, Mathew A. Crenson, and Brian L. Crissey.
"Public Decisionmaking in the Computer ERA," _Models in the Policy Process_.
New York: Russell Sage Foundation, 1976.

of testable outputs to inputs increases. However, this depends on the nature of the model and such considerations are more amenable to precise statistical analysis on a case-by-case basis.

An Alternative Paradigm for Validation

As indicated earlier, part of the confusion in discussions of validation for policy models result, one may argue, from misplaced views about the role of scientific proof. Most modelers have been trained in one or another field of the natural or social sciences and mathematics. They have an instilled tendency to judge an intellectual process by the tenets of modern science, which are dominantly empirical or logical positivist (i.e., any statement must be testable and a model must be validated against experience).

But is the paradigm of validation derived from scientific disciplines directly applicable to policy-oriented modeling? An argument can be made that policy-oriented modeling has a different objective than the use of models in science and must, therefore, be tested and evaluated against somewhat different criteria. Decisionmakers have to make decisions on problems for which science does not yet have definite answers. If the analyst holds strictly to that which is scientifically proven, he often will be in the position of telling his client that he can offer no advice whatsoever. Yet he often could offer some guidance based on partial information that arguably--and practically--may be better than no information at all.

It is fairly widely accepted that policy-oriented modeling should be viewed as part of a decision process, although this concept usually is not translated into precise criteria for evaluating models.6/ We can argue that statistical decision theories might provide alternative logical frameworks for the formulation and evaluation of models. In particular, several forms of decision theory, such as the Bayesian or subjective probability school, mini-maxers, etc., offer prescriptions for making decisions when information is incomplete or uncertain. Of course, there are well-known problems in social welfare theory in trying to impute values or preferences for society as a whole based on the type of preference that any single decisionmaker may favor.

There remains much to be accomplished in clarifying and refining the foundations of decision analysis and extending it to apply appropriately to public policy decisions. Nevertheless, one begins to perceive potential models for an alternative paradigm or at least a useful conceptual guide to further development of a more rigorous foundation for policy analysis. The key to these concepts is that they offer a systematic way of thinking about uncertain situations and alternative postulates for statistical inference. As an example, when viewing a set of data on health effects, a scientist seeks proof that it lies within a certain range

---

6/ Cf. Sisson, Roger L., "Introduction to Decision Models," in A Guide to Models in Governmental Planning to Operations, Gass, Saul, I., and Roger Sisson, (eds.) U.S. Environmental Protection Agency, 1974.

with high confidence, while the decision analyst seeks to extract a measure of the risks or probability that an effect will be in a certain range. When the scientist cannot prove his hypothesis, he punts; the decision analyst, on the fourth down, making the best use of whatever real information he has, has to play for the goal. The objectives are different and hence so are the statistical hypotheses that one seeks to establish from the data.

The key characteristic of any decision theory in relation to model validation purposes is not the decision rule, but the criteria for what constitutes information. This is, of course, an ancient problem in probability theory and the cause of much controversy. When probabilities are viewed as measures of ignorance, then the calculation of a priori probabilities always depends on what information we believe is known. A principal difference in schools of decision theory relates to the acceptance and treatment of judgmental and subjective information. Such information may not meet the test for scientific knowledge; but does it improve our guesswork when scientifically proven information is not available? Should we reject information that reduces uncertainty just because it is not in the form of a frequency distribution?

If we follow the general approach of decision analysis, we should view models as one of the means to estimate the probability of different outcomes resulting from identified policy options. A number of different uncertainties will apply in most cases, and we may classify them as follows:

- For problems that involve the future, there is uncertainty in future events that affect the outcome but are beyond the control of the decisionmaker.

- There is uncertainty about the nature of various biophysical, social, and economic processes affecting the outcome, which processes are represented by the model; these uncertainties relate to both facts that determine parameters in a model and the basic structure of the model.

The distinction between events and processes will depend in part on the boundaries of the system that one attempt to model.

Consider a simple situation where the model structure is well determined and only exogenous events and model parameters are uncertain. Then a decision analysis can be carried out by assigning probabilities to the events and parameters and using the model to determine the probabilities of all outcomes. In the subjective probability approach, the probabilities for events and parameters may be obtained from a combination of observed data and subjective judgement.

When the structure of the model itself is uncertain, the practical methods of decision analysis are more difficult to carry out. In this case, it would appear that one should postulate a spectrum of plausible models, calculate outcomes for each, and somehow weight the results according to the likelihood that each model may be correct.

In any approach, one will want to make use of the best information available, subject to cost-benefit considerations, in order to make the best decisions. One wants to validate models against real data where feasible. Thus, scientific knowledge will be useful and, by the very nature of its validation, preferable to subjective judgment.

A rigorous paradigm would be desirable to determine exactly how to blend subjective and objective information. Bayes original theorem was a beginning toward this problem; however, Bayesian decision analysis concepts may still be too vague to determine an exact approach. The problem becomes manifest when one moves from the subjective probability preferences of a single decisionmaker to the necessity for a more universally acceptable basis for public policy purposes. There presently does not appear to be a general principle or criterion to guide probability formation and the incorporation of subjective judgment. Added to that problem is the burden of identifying and weighting the spectrum of possible model structures. However, in spite of the difficulties attendant on uncertain models, we think there is some usefulness in decision analysis concepts as a general conceptual guideline for addressing the policy modeling problem.

To avoid misunderstanding, we should emphasize that the arguments we make here about limitations of natural and social sciences for predicting future events do not hinge on any assertions about fundamental limits of these sciences. Perhaps there are inherent limits on predictability, but such limits are not necessary to the argument here. It suffices to say that at the present state of scientific knowledge, many events and effects can not be predicted.

SOME TECHNICAL APPROACHES TO VALIDATION

Multimodel Testing

Using decision analysis with subjective probabilities as an analogue paradigm, one might defend the use of unvalidated models on the basis of making the best guesses about possible outcomes. But the analogue suggests that we then must also explore the spectrum of alternative plausible model structures, as well as alternative parameter values. To any modeler who has even considered making multiple runs of his model in order to reflect parameter ranges, the multimodel concept may be truly mind boggling! However, modelers often solve the parameter problems through combinations of theoretical and sensitivity analysis to select key parameters for variation. Similar devices probably are possible for attacking the multimodel problem.

One device that may assist the multimodel testing process is to separate large simulation models into components according to their potential for validation.[7] In the present approach, a component that is potentially validatable but not yet tested would be kept separated from judgmental components that are not capable of validation. The latter component might then be modified or substituted among

---

[7] Naylor and Finger have suggested a multistage verification process that could assist in the construction process. "Verification of Computer Simulation Models," _Management Sciences_, _14_, #2, October 1967.

ternative structures to achieve the multimodel spectrum while retaining the
validated portion of the model constant. This modular approach may render the
multimodel approach tractable. (However, the presence of feedbacks among the
components muddies separability.)

At the least, one might be able to perform sensitivity testing of the model to
see how key results are affected by the structural changes.

Reducing the Number of Variables

Other techniques are needed to assist the practical implementation of validation
and estimation of confidence levels.

As an example, where the output of a model cannot be validated directly, it is
often useful to be able to estimate the uncertainty in outputs due to given
uncertainties in input.

Recent thinking at the Department of Energy (DOE) has been directed toward various
means of grouping like variables so as to allow the statistician to operate on
fewer parameters. Probably the simplest idea would be to group variables on the
basis of the operation performed on them. All variables which were merely linear
coefficient relations of an earlier forecasted variable could be collapsed to
the control variable (as the variation would be a known function of the forcing
variable). The smaller the resultant number of variables, the easier would be
the application of statistical principles. It might actually be possible to
respecify these control variables by fiat or analytical choice, or the variables
agreed to be forcing could receive widespread review.

MODELS CAN BE PROVED CREDIBLE RATHER THAN VALIDATED

The Human Interface: Consensus is Not Validation

One practical form of validation is to be found in the judgment of the analysts
who have to use the forecasts. Recently, I heard this concept referred to as the
"laugh test." This feature was described as having a qualified analyst look at
the output of a run and see if he finds the results amusing. Such a test would
only exclude those runs which an analyst feels to give indeed impossible answers.

Perhaps a more acceptable approach, and much used in workday activities of a
policy shop, is to have a model run as a base or reference case, to change the
policy assumptions loaded into the model, and to rerun the model for comparative
purposes. This method places less emphasis on the absolute values of the forecasts,
and stresses the relative values resulting from a given set of policies. The
validation issue in this model use shifts from an emphasis on the forecasted values
per se to the reasonableness of the marginal change algorithm or the trends.[8]

Annual Environmental Assessment Report (AEAR). Washington, D.C.: U.S.
Department of Energy, Office of the Assistant Secretary for Environment, 1977,
for an example of the use of this technique.

But reasonableness is not validation. Probably the hardest thing for those who are new to policy analysis is to understand the frustrating fact that it probabl is not really possible to validate the data or results from policy models. Because no one is certain of the future, and because in our dynamic society the past is only a rough indicator of trends, the next best strategy has been to compare the results of a particular model with the results of other models or of surveys which purport to measure the same factors.9/ Although such comparisons clearly do not lead to much more than good feelings on the part of the modeler when his projections compare closely to someone or everyone else's, they are at least useful for pointing out where further analysis may be necessary. But thes comparisons do not shed much light on validity in cases where the results from various analysts are not in such tight agreement. Still this form of comparativ analysis does, at least, give the user some insight into how risky his results may be in the world of science and policy--close clustering would suggest less risk than does a scatter. Neither case should be mistaken for validity, however as all that is being tested in consensus. Except in a club, truth is not a commodity subject to vote.

Computers Permit Testing Alternative Measures of Reality

Our discussion thus far suggests a bleak and discouraging future for the complet validation of our models. It might also appear expedient to abandon attempts at validation, and look only at the usefulness of models for aiding the policymaker Nonetheless, numerous compelling arguments in support of modeling have appeared in the literature; so, even if we are unable to validate models, we do not have to apologize for using them. Instead, like good modelers, let us extrapolate so present trends.

Computers are coming into use in more and more sectors of our society. As exper ence increases and technology improves, the time cannot be far distant when the impact of computers will bring about societal changes as great as those from the industrial revolution. Computers are remarkable for their ability to store and organize and manipulate large amounts of information. It is inevitable that suc capability will be used to project into the future our present assumptions based on our present knowledge. After all, this is the way that the human mind tradi- tionally has made decisions, and certainly man will use in similar fashion the tool which extends his mental processes. The real question, therefore, is not whether he should place confidence in such devices, but how effective they are when compared to other available tools.

9/ A recent RANN study by Gianessi, L., and H. Peskin, <u>The Cost to Industries of Meeting the 1977 Provisions of the Water Pollution Control Amendments of 1972</u> Washington, D.C.: NBER, 1975. This is an example of such a validation attem In the end, it merely resulted in pointing out that different models yielded different results, but did not demonstrate the absolute or relative "correctn of anyone.

ow we can view the problem of validity from a different perspective. Models
ould be tested not only against reality but also against alternative methods
or representing reality; in particular, against mental models.

most no research in testing mental models has yet been carried out, and it
ould be very difficult. However, it can be argued that the standard validation
ests that some analysts and decisionmakers would require of models are signifi-
ntly more rigorous than those we require of alternative approaches, or of the
olicymaker himself. His decisions are subjected only to ex post facto validation
en he or his organization suffers the consequences of his invalid projections.

using either the mental model or the computer model to help make policy, a
ncerted effort on the part of the policymaker to act as if the model were true
nds to lead to self-fulfilling prophecy. The introduction of such purposeful
havior in the midst of any attempt to scientifically validate either the mental
the computer model raises serious doubts. However, the fact remains that the
mputer model is normally the only one required to prove its credibility in any
gorous fashion before it is used to guide policymaking.

WHAT USE WOULD VALIDATION BE TO THE POLICYMAKER?

es More Information Give Us Better Decisions?

ssibly the most difficult question relating to validation is not how it can be
ne, but what do the results mean and how should they be used. Let us assume,
spite of the reservations we have discussed earlier, that somehow or other
ocedures are established to perform the feat of validating model results. Let
further assume that the current trend toward models which are increasingly
mplex will continue. By complex, we mean not only are the models large in terms
numbers of variables, but that they are multidisciplinary: not simply economic,
vironmental, or engineering but containing variables from several fields. With-
t empirical evidence, we still hypothesize, on the basis of logic and experience
one, the inescapably questionable results of such model validations.

ere are many potential pitfalls in validation. Not only can one erroneously con-
ude that a model is valid when in fact it may not be, but one also can erroneously
nclude that it is not valid when it does have useful and valid properties. A
mple example of the latter is when a model predicts the wrong total value but is
le correctly to predict the increment due to a policy variable. The general
oblem is not merely one of applying the correct statistical tests to outputs, but
asking how the model will be used and judging it in that context.

e process of estimating error bounds is also prone to similar problems. As we
scribed earlier, theoretical estimates of errors propagated through large models
n seriously overestimate the resulting error bounds.

nce, both direct validation and error-estimating methods might show (if they
isted) that model forecasts had to be taken with large margins of error. In many
ses, these margins of error would be so apparently large that there would be some
estion as to the utility of the forecast at all. The problem is that inappropriate

validation procedures or conservative error bounds can lead to erroneous con-
clusions about the validity or suitability of a model for a given purpose.
Therein lies the other horn of the dilemma for policy models and validation.

Validation Tests: A Pandora's Box for Policymakers

The question boils down to how does one tell a policymaker about validation
tests? All kinds of issues bubble to the surface when this Pandora's box is
opened. For example, should one tell the policymaker the error expectations
associated with each analysis? If yes, how much should one then go into the
technical derivation of such information and the attendant error analysis?
Wouldn't subjecting analytical methods (particularly complex ones) to a rigorous
validating procedure bias the decisionmaker toward accepting nonrigorous
seat-of-the-pants estimates which are not presented to them with error
bands?

Actually, questions such as these are handled pragmatically. A good analysis
staff always tells the decisionmaker how much confidence they have in their
analysis, regardless of how this confidence was arrived at. Even if the con-
fidence is low, if a decision is to be made on some basis other than random
choice, the data and analysis—regardless of how suspect—will normally form the
basis for choice. Although decisionmakers do sometimes use analysis to leap to
a preconceived conclusion, such bias cannot possibly be ascribed to the incredibly
large number of decisions these people have to make. Assuming some rationality,
the decisionmaker has to trust the inputs from staff, based on whatever information
there is available at that time. Although the quality of the data may impact the
form of the decision, it will seldom determine whether a decision is made. Be-
cause no decisionmaker is omnipotent, their decisions are usually forced by action
taken elsewhere. The option of not making a decision because of a lack of high
quality information is a luxury that cannot be afforded.

Conclusions

In summary, it is clear that from a practical viewpoint, we have had little to show
for rigorous, total validation. Although our technical backgrounds would lead us
to applaud the effort to discover and perfect such procedures, we hold little hope
for their being found, especially for large-scale complex models.

Validation of certain components of policy models may be useful, however, to the
extent that it treats such component models in their proper context. Thus, valida-
tion has a place, but it is unlikely to be a comprehensive tool.

We have raised more fundamental questions about the appropriateness of validation
Where it is not appropriate, the exercise of attempting it may be misleading and
even harmful to modeling and policy analysis. If we can identify more clearly
where and in what sense models can validated, perhaps we can improve their utility
and acceptance as well as avoiding a great deal of needless argument among modelers

Although there is some cause for despair about the analytical postulations we have presented, the practical utility of the results should not appear startling. No practicing policy analyst or experienced decisionmaker ever takes as gospel any analysis, regardless of how derived. Our institutional structures are designed so that major decisions can be revisited and revised as more information becomes available or as experience changes the perceptions and goals. Finding out analytically that we are not very sure of what will really happen if a particular policy choice is made is not news to the experienced. What it may do, though, is allow those dastardly policymakers who have a preconceived notion·to ignore the results of analytical investigations on the basis of inappropriate validation findings to turn to their own subjective notions. This is the very event that early modelers were trying to avoid with such procedures as systems analysis: an interesting, and unvalidated, speculation with counterintuitive results.

DISCUSSION

Mr. Joel (NBS): First of all, Dr. House mentioned five characteristics that he wanted to consider, and when you said "size," I thought at first you were talking about physical size of the subject and it turns out you were talking about the number of equations in the model. I suggest that there is a sixth characteristic then, and that's the physical size of the model subject. Because if the size is sufficiently big or sufficiently small, that's when you run into problems of direct observation for validating purposes. Point number one.

Point number two about forecasting the past. Most people when they are trying to validate forecast models don't really forecast the past. What they try to do is see if they can reproduce a time series of some sort. There's a sort of slightly different notion, that is the invertability of a model. To consider the current set of states of something. Then to go back, using that set, and have an inverse model to show what some of the--to try to predict the range of the inputs that could have led to that sort of thing. It gives you a much better subjective feeling for what the forecasting power, the divination, the power that's built into a model would be like. It still wouldn't give you a way to assign a number to it. That is to say, this model has 42 percent accuracy, but it will give you a better feel.

On the idea of a utility of a model in aiding decisions. That's a very pretty phrase, but I noticed that you didn't really have any kind of measure of that sort of thing to propose. And I suggest that it's very hard. I don't even know if there is a way to define what you mean by that in quantitative terms.

Just a couple of semi-educated remarks about some things. I do know that in models of hydrological systems, there has been some interest recently in using information theoretic measures of accuracy, of the amount of information in a time sequence, and it may be that that sort of tool would be useful in validation of models.

The other thing is, we talked about defining consensus. There is a technique which is widely ridiculed by serious scientists. It's used by a lot of charlatans. I suggest that the Delphi technique is not to be considered as a method of finding truth, but as a means for defining consensus on various kinds of questions, it might be very useful--especially in defining a yardstick for validating models.

One other thing. I would prefer to talk about the objects of models rather than models themselves. In using your hierarchical approach to defining the relative difficulty of validation, what you really want to talk about is the subject systems themselves and which types of systems are more amenable to certain kinds of observations and statements.

Then you will avoid this pervasive policy of regarding the model as the "Ding an sich" that really has some meaning when it really was supposed to have been a representation of something else to start with. It's a sort of a bad habit, but people do it.

I noticed you carefully avoided talking about policy models and science models. And two of the people this morning weren't so meticulous. And as a consequence, instead of talking about using models for policy purposes and using models for research purposes, they started to use phrases like "policy models" and "science models"; after a while it's not hard to believe that such objects actually exist, with those descriptions. And I don't believe there is any such thing as a policy model or a science model.

Dr. Ball: I think he made a number of good points. I just have a comment on the questions of trying to use models backward. The problem is, how much of the world is really reversible or one to one in that sense?

Mr. Joel: It gives you a good mirror of what the difficulty was of representing it in the first place. Because you made certain comments and assumptions in making a representation to the relationship of objects in the universe. You may be rash about that. You'll find out a lot quicker because of the sense of it.

Dr. Ball: I think your comments about looking at the type of the system rather than the type of model is a good one, well taken. I would like to explore that further.

I think Delphi would represent one of the possible solutions to the problem of how you incorporate judgmental information. And I guess the point that I was really getting at was that we haven't yet learned quite how to take decision theory over completely, rigorously, into the group process.

Mr. Joel: I shouldn't interrupt, but I didn't say Delphi should be used exclusively to define the consensus, not really.

Dr. Ball: Well, it's also possible that one would want it. I think people have suggested--maybe you didn't--people have suggested Delphi as a process for arriving at judgmental information. The only point I would want to make is that we don't have good criteria for what is the best process to arrive at that judgmental information once we get to the multiple decision-maker process. I don't think we're ever going to arrive at a complete rigorous theory there, but a little bit more thought in that regard might suggest, at least qualitatively, some criteria for incorporating expert judgment in a more systematic way.

Dr. Glassey (DOE): Let me return to a point of philosophy a minute. Being an empiricist, I guess, myself, I was struck by your apparent suggestion that modelers might adopt another paradigm other than the scientific

paradigm. Now, I think I understand that policy makers often do not use the scientific paradigm as a means of arriving at policy. But, are you suggesting, in fact, that model makers abandon the scientific paradigm, and embrace the notion of judgmentally based recommendations as a means of operating as modelers?

Dr. Ball: I'm not suggesting that we abandon the scientific paradigm. What I'm suggesting is that the scientific paradigm only takes you so far. We are dealing with problems where science has nothing to say. Wherever science has something to say, I would immediately agree that it is the best way to arrive at a conclusion; where science can contribute information and validate it scientifically, that is, by the very nature of its validation, it is the best information we can get and I would want to use it.

The problem is that science gets to certain problems and then it just punts. It can't help us, and we still have a decision to make. Therefore, we need to say how can we go beyond that and make the best guess, even though we can't scientifically prove it. I'm saying that we need a paradigm to fill in a gap that isn't being filled by the scientific paradigm.

Dr. Glassey: Let me suggest that that's not the role of the modeler. The role of the modeler when he reaches the end of the limits of his science, at that point should confess that he has come to the end of his knowledge as a modeler, and then fall silent. I don't think that there are modelers of any particular expertise in the area of judgment about policy beyond that which we derive from a scientific method.

Dr. House: Roger, would you do me a favor and I won't have to answer that. Would you tell me the truth areas, and then I'll quit after that. I think there are--except for--I don't even know how to handle some of that. Let me try talking about a model in a slightly different way.

I think that the attack that both Dick and I have taken on validation is almost exactly aligned to what you're talking about. Let me handle it as a caricature because it might work better that way.

The field of attempting to talk of validation only in terms of aping, either a past or another set of principles, probably for a whole class of models isn't really possible, because we just never know, or at least we don't have enough information now to know what the truth really is.

But there is a whole useful part of--for example, when you were talking about the hydrological models, I frankly would put that into a class where I would agree with Roger. I think we can, after looking at stream flows over a long period of time, have an empirical data base on which we can build a model and use that for forecasting of a particular type. I think that's something that a model does very, very well.

Our great trouble, for example, except in playing "what if" games with taking models built in the econometric area for a whole series of things, even of a single firm over a period of time, is in trying to talk in terms of anything but the grossest sample, given the fact that they follow

everything that they did in the past. If you do this, this is the kind
of impact that you'd get out of it. But hardly in terms of validation,
in terms of a scientific field.

Now, we used to use terms like "veracity," "verisimilitude,"--all of those
wonderful words. I think everybody ran through a dictionary and tried
to find something that sounded the same as validation.

But, what these were was slightly more than a warm tummy feeling. And,
it seems to me to be perfectly legitimate for anybody that uses a model
or buys results from the model, to at least make sure that the model does
what the people or person who built it says it does. You know, when you
punch on it in one place, it pops at the same level that you'd expected
it to pop on the other side. But to talk about that in terms of valida-
tion is saying that that thing that pops out on the other end has anything
related to truth, reality, or anything else on the other end; I can't
go for it.

However, I don't think you ought to make the mistake of throwing the
baby out with the bath water at the other end by saying that because I
can't assign a number to truth on the other end of it--that it's garbage.
I mean that for some reason or other, you have to say that that's the
best I can do.

There is a set of decisions that have to be done; you make them everyday,
you do. And you make them on the best set of information that you have,
and the only thing that a model can do for you, I think in some of this
area, is specify the assumptions that you made for it. And maybe that
makes a better decision.

   Mr. Woods (GAO): And I have a question to ask, seeing how it's sort
of going back and forth, but I get the feeling that validation, what we
call this generic term "validation," there almost seem to be two categories
in it.

Number one is that now you're talking in terms of a human being rather than
in terms of a model; does he know what he's talking about? Number two,
does he tell the truth?

In a sense that I can classify, for instance, you are a well-trained opera-
tions researcher or physicist; but I also know that if your boss asked
for crap, you'll give it. In other words, there is a question between
validity in a mechanistic sense and credibility in the sense that I want
to use it.

Now, I get the feeling that, somehow or other, that the two become garbled
and I'm still trying to understand. Because everybody might suddenly
say that a particular model has a tremendous amount of mechanical credibility.
But, when you look at how it is used, you cannot divorce the two of them.
And the question is, can you--I get the feeling that what you're talking
about in terms of validity is more of a mechanistic. In other words, I
am making sure that, in fact, the guy meets all of the licensing requirements
for being an M.D. or Ph.D., or are you talking in terms of credibility
as well?

Dr. Ball: I wasn't talking in terms of credibility. Validity, in the sense that we are using it here, has to do with the concept of comparing things against reality. Validation implies, in the sense that we're using it, that you are comparing the output directly against observed reality and seeing whether it's correct. Validity is often used in other senses as well, including your sense of credibility. But, I don't think we're talking about it in that sense.

I would suggest--I don't know whether this is what you're driving at-- but if you used this multi-model or component concept that I was talking about, it would help you to maintain the credibility aspect purely by making what you do more open and transparent. Mainly, if you keep the parts of the model that are judgmental, the parts of the input assumptions that are judgmental, clear of the other mechanistic parts of the model and lay them out for people as clearly as possible, that's the only way I know of to be credible in that way. You'd simply have to open it up and let people take their choice. And if you can make it clear enough so that people can see what you're doing, they will have to make their own decision. Then your own credibility is not so much questioned in the process.

Mr. Woods: I guess the thing is that there has been a tremendous amount of discussion. I'm looking at the calendar for the future, both today and tomorrow on whether there seems to be a focus on how do we mechanically go about back-fitting, etc., etc.

I get the feeling that credibility is the most critical thing, and the question of the validity--whether or not it meets all the mechanistic things--is sort of secondary. So the first thing that you must do when you set up the system is to establish credibility rather than this mechanistic validation process.

Dr. Ball: I wouldn't maintain that. No. I would maintain that credibility is a meaningless thing, particularly in Government, in that you can't even attempt--. Openness is the only possible way that you can deal with that part--.

Dr. House: If you'd like to use the word "validation," that's what I really meant by "veracity" or "verisimilitude." For scientific validation, and I think that is what we were talking about here, that's a necessary but not a sufficient condition. Almost by definition, you'd have to have that in order to get comparative analysis to say that it turned out that way.

I guess the major split, and this is probably a good way to split it as almost any other, models at least should pass the first test. I'm abso- lutely certain that a large number of the models that I have looked at over the many years don't pass that first test. I mean, they don't do what they say they do and they don't do it except under the most rigorous conditions. I've seen more model outputs produced by a typewriter than I have by a computer. And, so, that's at least a necessary condition. Now the second condition of comparative analysis is some sort of objection to truth. I'm just saying that there are places where that falls apart.

Now, Dick and I have talked about this attempt to partition the types of models so that they might handle that. Some models fall very easily into it and others don't and maybe what you'd like to do is separate them and the like. But there are just some that can't fall into the second case because we just don't know enough to put them there. And, by the way, the classification we had out there, some type of models weren't designed to do that. I think your optimization work may be one. I quit. Thank you.

THIRD PARTY MODEL ASSESSMENT: A SPONSOR'S PERSPECTIVE

Richard Richels
Electric Power Research Institute

INTRODUCTION

The electric power industry has long been a sophisticated builder and user
of models in planning capacity expansion, and in scheduling existing gener-
ation capacity to satisfy customers' energy demands at minimum cost. More
recently the industry has provided support for developing large-scale mod-
els that encompass the interactions between the electricity sector of the
economy and the rest of the energy-economic system. In addition, there
are a number of important models which, although not used directly by the
electric power industry, play a role in determining public policies that
affect the industry. Sponsors of these more general models include pri-
vate foundations, the National Science Foundation and government agencies.

A sampling of models relevant to the electric utility industry is as fol-
lows: The Baughman-Joskow Regionalized Electricity Model, the Wharton
Macroeconomic Energy Model, the Hudson-Jorgenson Macroeconomic Energy Mod-
el, the Gulf-SRI Energy System Model, the Brookhaven Energy System Opti-
mization Model, the FEA Project Independence Evaluation System (PIES), the
ETA-MACRO Model and the ICF Coal and Electric Utilities Model. Each of these
models includes an explicit representation of the electric power sector and,
to varying degrees, all are being used in technology assessment and/or
policy analysis relevant to the electric power industry. It is important
for the electric utility industry to be certain that such models accurately
represent the "real" world. This is the basic rationale behind the Electric
Power Research Institute sponsoring the Model Verification and Assessment
Project (RP 1015) at the MIT Energy Laboratory (1).

The model verification and assessment project was initiated on a trial basis to test the practicality and usefulness of third-party model analysis. The Model Assessment Laboratory has three objectives. It is intended to:

1. provide model users with evaluative information and understanding essential for the intelligent use of models;

2. give model builders feedback signals helpful in correcting and improving the models; and

3. promote the development of state-of-the-art model assessment.

These three objectives are considered central to the health and usefulness of the energy modeling field and to the development of an infrastructure of supporting services and criticism.

In the past EPRI has turned to individual investigators for independent assessments of selected energy models. For example, in an analysis of the Brookhaven Models, the assessors were asked to evaluate the electric utility sector of the models and then to modify, extend and refine the models in relation to the existing "state-of-the-art" (2). Such assessments are typically "one-shot" efforts with little thought to developing a set of assessment criteria or establishing an assessment methodology which could form the foundation of future assessments. By consolidating the assessment function under a single organization, it is hoped that "life" and "continuity" can be brought to the assessment process and that the understanding and insights gained from past assessments could benefit future work.

Composition of the Assessment Laboratory

The model assessment facility employs two types of researchers, those that

form the infrastructure of the laboratory and provide the continuity among assessments and those that come on board for a particular assessment because of their expertise regarding certain aspects of the model undergoing assessment. The first category, the "model analyzer", represents a new type of researcher, "a highly skilled professional and astute practitioner of the art and science of third party model analysis" (3). Ideally he will be experienced both as a model builder and a model user but occupy a "middle position" while involved in the assessment process. The assessment group's reputation for fairness and objectivity depends heavily upon adherence to this condition. The model assessment process must provide for frequent interactions between the assessment and modeler groups. Modelers may feel uncomfortable about discussing model deficiencies, if they perceive the assessors as competitors or potential users.

Energy models employ the analytic methods of a variety of disciplines. A single modeling system may incorporate the techniques of several disciplines. If the laboratory is to maintain a capability for assessing a wide range of energy models, it must also have a strong resource base of researchers that can be drawn upon for specific assessments. Included in this group will be experts in mathematical programming, econometrics and related methods of statistical analysis as well as experts in various aspects of the electric utility industry and the energy sector in general. The composition of the assessment team at any given point will depend upon the characteristics of the model undergoing assessment.

Approaches to Assessment

The model assessment lab provides two types of assessments: (a) overview

assessments of selected models in which EPRI has an interest to determine particular strengths and weaknesses for dealing with a specified class of technology assessment, and/or policy analysis. Such overview assessments are appropriate for determining models for which a more detailed assessment is required. (b) more detailed critical assessments for models which EPRI intends to use extensively. Such assessments provide for indepth analysis of model formulation, of data development and integrity, and of appropriateness of statistical estimation techniques. They also provide replications of statistical estimation and simulation results, and sensitivity studies of critical points. A key distinction between an overview assessment and an indepth assessment is that the latter includes the complete assimilation of the model on the model assessment facilities' computer system.

The REM Assessment

The first year of the MIT effort included both overview and indepth assessments of the Regionalized Electricity Model (REM), designed to determine the relative advantages of each type of analysis (4). REM was constructed to analyze policy issues affecting electricity producers, consumers, regulators, and equipment vendors; issues such as peak-load pricing, inclusion of work-in-progress in the rate base, and the costs of environmental standards. Unlike previous electricity sector models, REM provides for simultaneous linkage of supply, demand, pricing, and financial behavior in a single integrated framework merging economic-engineering, behavioral, financial, and econometric models.

REM was used in 1976 to examine the future of the U.S. nuclear industry (5). The study concluded that the industry and the Atomic Energy Commission were

174

substantially overestimating the growth of nuclear power through the end of the 20th century, and it raised serious questions as to the future financial viability of the nuclear equipment manufacturers.

The assessment was divided into the same three categories as the model: demand, supply, and financial/regulatory, with teams of researchers assigned to each. For the overview, the REM computer code served as the source document on the model's operations. The first step in the overview assessment was a line-by-line analysis of the code in order to verify the implementation of the research design.

The major focus of the overview assessment was face validation. Here the assessors subjectively evaluate the degree to which the model corresponds to their perceptions about the actual phenomona being modeled. The assessment group relied on both its own background in electric utility modeling and the experience of an advisory panel drawn from the utility industry. The product of the overview assessment was a general statement on the overall structure and implementation of each submodel and a list of features requiring additional investigation in the indepth assessment.

Even with the substantial resources available to the project, care had to be taken in selecting areas for indepth analysis. The transmission and distribution and nuclear components of the supply submodel were excluded, for example, not because they were considered unimportant, but because it seemed more productive to concentrate on the other three components of the supply submodel.

Especially valuable to potential users are the assessment's conclusions concerning the model's applicability to specific policy issues. A list of policy applications - present and potential - is given in Table 1.

The report, where possible, reached definite conclusions concerning the usefulness of the model for policy analysis, identified policy parameters, and noted the sensitivity of results to changes in input assumptions. It also specified how the model might be modified to enable it to treat issues for which it was not specifically designed.

We now turn to a discussion of some of the major problems and lessons from the REM assessment.

The Moving Target Problem

By the conclusion of the REM assessment, the version of the model which occupied the attention of the assessors, in a sense no longer existed. REM is an active research tool. As the model is brought to new applications it is modified and improved. To a certain extent an active model can be thought of as a "living" model - constantly in a state of evolution. This presents certain problems for the assessment process. It is essential that a single version of the model be "frozen" in time for purposes of assessment. Otherwise the assessment laboratory will be, in a sense, "shooting at a moving target." Yet if assessments are to remain timely, attention must be paid to the dynamic nature of the modeling process.

One approach to dealing with the problem is for the assessment laboratory to undertake periodic audits of REM. This might involve visiting with the

Table 1

Potential Policy Applications Considered in Assessing REM

---

1. *Changes in Factors Affecting Electricity Demand Growth Paths*
   - Economic/demographic trends
   - Conservation policies

2. *Load Management*
   - Peak load pricing*
   - Cogeneration
   - Seasonal pricing

3. *Impacts of Changes in Cost Factors*
   - Capital costs for new plants
   - Fuel prices*
   - Wage rates
   - Taxes (possibly a Btu tax)

4. *Changes in Resource Supply Conditions*
   - Resource constraints*
   - Increasing cost supply schedules

5. *Costs of Financing**

6. *Industry Responses to Capital "Shortage"*
   - State financing*
   - Less capital-intensive technologies
   - Reduce growth
   - Reduction in plant reserve margin

7. *Regulatory Policies*
   - Regulated rate of return*
   - Inclusion of work in progress in rate base*
   - Exclusion of noneconomic plants from rate base*
   - Regulatory lag

8. *Alternative Lead Times for Capacity Expansion**

9. *Environmental Constraints*
   - Siting restrictions
   - Capital equipment requirements*
   - Increased operating costs*

10. *Technology Assessment*
    - Advanced generation technologies: Centralized and distributed conventional and nonconventional cogeneration fuel conversion
    - Nuclear: Non-LWR, breeder, etc.
    - Storage
    - T and D

---

*Applications for which examples already exist in the REM documentation.

Source: (1)

modelers, reviewing recent changes, and having the modelers produce a series of computer runs that could be used for validation. The assessors could also review the current version of the computer code for purposes of verification and assess the quality of the revised documentation. The advantage of this approach is that it circumvents the costly and time consuming process of assimilating the new version of the model on the assessment laboratories computer facilities. Supplements could periodically be appended to the initial report to insure that the assessment evolves hand in hand with the model.

Comparative Versus Individual Assessments

The assessment of an electricity model such as REM requires a team of researchers with a wide range of expertise. Individuals are required with backgrounds in econometrics, statistical analysis, computer programming as well as experts with experience in various aspects of the electric utility industry and the energy sector in general. Clearly, once such a team is assembled, the assessment laboratory is in a position not only to assess REM but other electricity models of interest to EPRI as well. This raises the issue of comparative assessments and the extent to which the models being assessed should be compared with similar models.

There are reasons for comparative assessments besides efficiency. Modelers have noted that assessing only one of several models appropriate to a particular class of policy issues can unjustifiably discredit that model in the eyes of potential users. Many of the weaknesses attributed to REM are common to most electricity models. The "all-or-nothing" character of REM's capacity expansion and generation mix decisions is an example of a problem generic to

178

most modeling methodologies. When compared with all other related models, a perceived weakness in REM may turn out to be a comparative advantage. An assessment report which focuses on the relative strengths and weakness of competing models offers more to potential users than is possible through individual assessment.

Relations Between the Model Assessment Group, the Modelers, and the Model Assessment Sponsor

Once the objectivity of the assessment group has been established, a protocol must be developed to allow for the interactions between the assessment and modeler groups. For the REM assessment, a detailed statement of work between EPRI and the MIT model assessment group laid out the schedule of activities, deliverables, and financial resources to be devoted to the project. Although it was realized that the cooperation of the modelers was essential to the assessment process, no formal contract or statement of understanding existed between the sponsor and the modelers. This resulted in the modelers spending significant unreimbursed resources in time and materials to participate in project review meetings, to review and comment on draft materials and to prepare formal comments on the REM assessment. It is now clear that in future assessments, arrangements should be made between the modelers and sponsors regarding the terms and conditions for modeler participation.

Conclusions and Recommendations

The project is a major step toward developing effective procedures for the independent evaluation of energy models. Independent model assessment is a critical element in making models more accessible and useful in the areas of technology assessment and policy analysis. Such "third-party" assess-

ments identify the weak points of a model's theoretical structure, empirical techniques, and implementation procedures. This information can then be used to employ the existing model most fruitfully and to develop future models.

Rather than being an adversary proceeding, independent assessment is most effective when undertaken as a cooperative venture with support from and dialogue with the modeler. Independent assessment can increase confidence not only in a particular model but also in the credibility of the model developer, who has an important expert role to play in technology assessment and policy analysis.

During its first year, MIT's Model Assessment Lab undertook overview assessments of the Baughman-Joskow Regionalized Electricity Model (REM) and the Wharton Macroeconomic Energy Model, and an indepth assessment of REM. The research has been successful and has played an important role in shaping EPRI's use of these models. Third-party verification and assessment has enabled EPRI to understand better the strengths and deficiencies of the models and to improve their use in planning research and making actual studies.

At the present time, an indepth assessment is being made of the ICF Coal and Electric Utilities Model. This model is important to the electric utility industry because it is being used by the Department of Energy and the Environmental Protection Agency to assess the effects of alternative new source performance standards for coal-burning power plants. The prototype venture at MIT has proven the merit of the idea of a model assessment facility and justified the plan to institute such an operation on a continuing long-term basis.

Footnotes

1. Massachusetts Institute of Technology, Energy Laboratory. 1978.
    Independent Assessment of Energy Policy Models: Two Case Stu-
    dies. Report No. 78-011. Cambridge: MIT.

2. Systems Control, Inc. 1978. _Applicability of Brookhaven National_
    Laboratory's Energy Models to Electric Utility R & D Planning.
    EPRI EA-807. Palo Alto, California: Electric Power Research
    Institute.

3. Greenberger, M., Crenson, M. A., Crissey, B. L. 1976. _Models in_
    the Policy Process. New York: Russell Sage Foundation.

4. The discussion of the REM assessment is abstracted from Greenberger, M.,
    and Richels, R. 1979. Assessing Energy Policy Models: Current
    State and Future Directions. _Annual Review of Energy_. 4.
    Palo Alto, California: Annual Reviews, Inc.

5. Joskow, P. L., Baughman, M. L. 1976. The Future of the U. S.
    Electric Utility Industry. _The Bell Journal of Economics_.
    7: 3-32.

An Approach to Independent Model Assessment

David T. Kresge*

The M.I.T. Energy Laboratory has recently completed a study, funded by EPRI, dealing with the independent assessment of models used for energy policy analysis. The principal objectives of the project were to:

- Provide assessments of two important energy systems models, the Baughman-Joskow Regionalized Electricity Model and the Wharton Annual Energy Model;
- Analyze these case studies to identify key organizational and procedural issues which must be addressed in the assessment process;
- Develop a framework for better understanding the general approaches to and objectives of energy model assessment.

This paper draws on the experience gained through that assessment project to (1) present a general outline of the approaches to model assessment; (2) report on some of the specific lessons learned; and (3) make some suggestions for improvements in future assessment activities.

Approaches to Model Assessment: A Working Hypothesis

The framework for energy model assessment which we propose as a working (and we hope workable) hypothesis contains four principal elements: (1) review of literature; (2) overview model assessment; (3) independent audit, and (4) in-depth assessment. Although these elements represent four distinct approaches to model assessment, they are most appropriately viewed as the stages in a comprehensive model assessment process. The approaches are interactive and complementary, and should not be viewed as mutually exclusive alternatives.

A summary of the content and relationships among the approaches to energy model assessment is given in Figure 1. The assessment process must begin, of course, with an operational version of the energy model to be assessed. For a reasonably mature model, the available documentation should include at a minimum: a concise statement of the model's conceptual structure; a description of the procedures by which the model was empirically implemented (including a discussion of the underlying data bases); and a discussion of the results obtained when the model was applied to an analysis of the policy issues for which it was designed.

David T. Kresge, Associate Director, Joint Center for Urban Studies of the Massachusetts Institute of Technology and Harvard University, Cambridge, Massachusetts

FIGURE 1

APPROACHES TO ENERGY MODEL ASSESSMENT
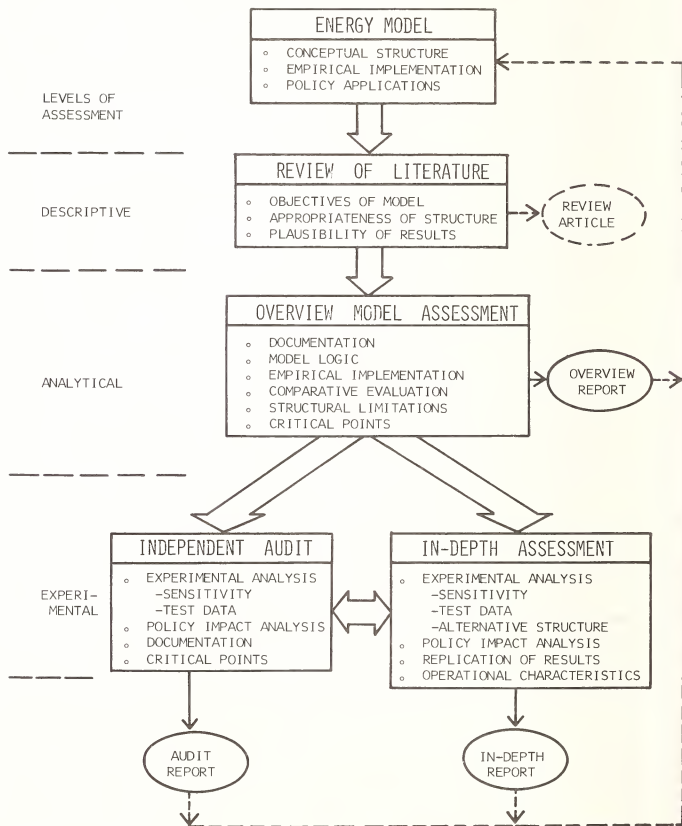
For a model still in the development stage, the available documentation would be expected to be more rudimentary, but assessment of such a model is still both feasible and desirable. Indeed, in our opinion, carrying out assessment in parallel with model development is one of the most promising avenues for improving the credibility and reliability of energy policy models. The documentation available to such an assessment effort might take the form of working papers which would then be supplemented through direct discussions with the model builders. Information relating to policy implications might need to be generated as part of the assessment process. Further observations concerning the procedures for assessing a model still in the development state will be incorporated in the later discussions of assessment approaches.

.2. Review of Literature

An evaluation of a model's structure and characteristics which relies solely on published materials dealing with the model is what we term a "review of literature." Such a review, which is an essential first step in any assessment process, brings together and summarizes the available published information describing the model's objectives, structure, and principal results. As indicated in figure 1, this information is often presented in the form of a review article.

The author of such an article will generally try to make some evaluation of the appropriateness of the model's structure for dealing with the policy issues on which it is focused. Often the plausibility of the results will also be judged through comparison with the results produced by other related pieces of analysis. Clearly the evaluative component of the review article depends critically on the expertise of the author and on the completeness of the documentation.

A review of literature is useful but, in our opinion, is essentially a descriptive procedure. Published materials discussing the model's structure, implementation, and applications are generally so highly condensed that they do not provide adequate basis for making well-informed judgments concerning the model's validity. A review article is primarily useful to a potential model user in providing a description of what the model is intended to do and of the methodology used to achieve the stated objectives.

3. Overview Model Assessment

Overview model assessment, the next stage in the general assessment process, goes beyond literature review by turning to the underlying (and generally unpublished) technical documentation. An overview assessment is primarily an analytical evaluation of the model's properties. An overview model assessment report can be expected to contain three major types of information: (1) an evaluation of the empirical content of the model, perhaps with comparison to other

empirical studies of similar components; (2) A discussion of the limitations on the model's applicability due to its basic structure; and (3) Identification of the critical points and issues in the model's structure, empirical content, and applications which require further experimental analysis. In our experience, the most important element of such documentation is the computer code used to implement the model. In contrast with the published material, computer code has the very desirable property of leaving nothing to the imagination since every operation must be stated explicitly and unambiguously. Unfortunately, the interpretation of computer code is often a very difficult task which may well demand an even higher level of programming skill than was required to build the model in the first place. It is our judgment, however, that analysis of the computer code make such a significant contribution to an overview assessment that it is in general, worth the cost entailed.

An overview assessment should include a comprehensive list of policy applications that might be considered by potential users with limited knowledge of the model. Detailed analysis of a model's structure will often show that there are seemingly plausible applications for which the model is actually ill-suited. An overview report should point out these inherent limitations on the model's applicability. This information can assist potential users even when the proposed applications are ones that the modelers have never suggested. The list of potential applications also helps define the context within which the assessment was carried out.

One of the most important features of the overview report is an identification of the model's major "critical points." A critical point for our purposes is defined as an element of the model about which other experts might raise questions and which is expected to have a significant influence on the model's behavior.* A listing of the model's critical points can often serve as a concise summary of the principal findings of the overview assessment. Developing such a list and providing reasons for each item included in the list should be a primary objective of the overview report.

Although an overview report should be able to identify a model critical points, it will only rarely be able to pass judgment on the adequacy of a model's treatment of them. A critical point is, by definition, an issue on which reasonable, well-informed analysts disagree. It is, therefore, an issue which often cannot be settled by the analytical treatment in an overview assessment. Thus, an overview report is actually an interim document in which many questions are raised but only a few are answered. Further assessment of

---

*The term "critical point" is very closely related to the concept of "contention point" and "critical contention point" as used by Crissey and others. There are some minor differences in the two concepts so we chose to use a different phrase to avoid confusion.

the model's validity requires the acquisition and analysis of experimental data. Such data are essential if the assessment process is to produce substantive conclusions concerning the model's critical points. Although the overview assessment is generally not able to produce such conclusions, it does, by systematically identifying the critical points, provide a sound basis for the next stage of the assessment process.

1.4. Independent Audit

An independent audit uses data derived from experiments run with the model to evaluate the model's validity, applicability, and performance. The experiments are designed by the assessors but are implemented by the modelers, with the proviso, however, that a member of the assessment group be present as an observer when the experiments are run. It is our view that this "looking over the shoulder" element of the procedure is essential to the accurate interpretation of the results produced by the experiment. The outcome of an experiment is frequently influenced subtly but critically by the way in which it is implemented.

An audit report should use the experimental data together with the analytical material developed in previous stages of the assessment process to provide an evaluation of the model's validity in as many key areas as is feasible. In particular, the report should focus on the model's behavior with regard to its major critical points. The audit report should also provide information on the quality of the available documentation. It is our experience that when a model's behavior differs from what was expected, it is often due to incorrect or unclear documentation. There are also instances in which the documentation is correct, but errors in implementation prevent the model from doing what it is supposed to do. In either case, the report should point out such discrepancies, both to potential users and to the modelers.

It should be noted that an independent audit will generally not be able to make definitive judgments concerning all critical points that have been identified. Some points can be investigated only through structural analysis too complex to be handled within the audit approach. On other points, the audit may be able to show that the model behaves in ways that seem inappropriate, but will not be able to show why the model behaves as it does. In these instances, the experimental data generated in the audit are able to push the analysis further than was possible in the overview assessment, but it is not sufficient to make a complete, definitive assessment. For the critical points requiring this more complex type of analysis, it is necessary to proceed to an in-depth assessment.

1.5. In-depth Assessment

An in-depth assessment, like an independent audit, relies heavily on the analysis of experimental data. The difference is that the in-depth assessment generates some or all of the data through direct, hands-on operation of the model. Direct operation makes it feasible to carry out much more complex tests, particularly when the tests involve making modifications in the model structure rather than simply changing model parameters or data. Another rationale for the procedure is that the closer one gets to the operation of a model, the more likely one is to identify errors and discrepancies between implementation and documentation.

As indicated in Figure 1, an in-depth assessment could conceivably be undertaken either immediately subsequent to an overview assessment or after an independent audit had first been completed. Because an in-depth assessment is such a substantial undertaking, it is our view that it is usually most efficient to first conduct exploratory analysis through an independent audit. The audit will also allow the assessment group to gain familiarity with the model by working with the modelers before attempting to run the model themselves. Furthermore, in some instances the results of the audit may be so conclusive that it is decided that there is no need to proceed with the in-depth assessment.

Since any major energy policy model will undergo a virtually continuous process of change, the in-depth report may also be able to contribute to later modifications or extensions in the modeling framework. Unless the modifications are so extensive that they result in a completely new model, the appropriate way to update the assessment would be to use it as the starting point for an independent audit. This is why Figure 1 shows an arrow leading from in-depth assessment to audit as well as from audit to in-depth assessment. With the in-depth assessment as the base, the update audit would, of course, focus on those features of the model which had been modified. With so much previous materials and expertise to draw upon, the cost of such an audit would be quite modest and would provide a very efficient means for updating the assessment reports as new versions of the model are developed.

2. In-Depth Assessment: The Baughman-Joskow Regionalized Electricity Model

The assessment procedures applied to the Baughman-Joskow Regionalized Electricity Model (REM) are outlined in Figure 2, where the shaded portions indicate the steps included in the REM assessment. After completing a review of the relevant literature, we concluded that it did not provide an adequate basis for a legitimate assessment, so a review article was not issued. Our review of the literature further convinced us that the overview assessment could not be based solely on published materials but would have to rely heavily on analysis of the computer code. Even though we regard the quality of the REM documentation as above average relative to comparable models, we found a number of instances in which the documentation was incomplete or inconsistent with the actual code.

After completing the overview analysis, we did issue an overview assessment report. This report was, however, treated as an internal working document and was made available only to the assessment group, the sponsors, and the modelers. After a review meeting in which Professor Martin Baughman participated, it was agreed that all further analysis would be based on an updated version of the model in which some of the errors identified in the overview report had been corrected. An in-depth assessment was undertaken immediately upon completing the overview assessment. An independent audit was not included in the assessment procedures applied to REM.

To help structure our assessment of REM, we compiled a list of potential policy applications. This list, which is shown in Table 1, included both the issues to which the model had already been applied and those issues for which, in our opinion, further applications might be considered. There is no implication that the model builders themselves would suggest that REM is appropriate for all the applications shown in the table. We tried to make the list wide-ranging in order to make our observations useful to potential users who were not familiar with the model's properties. The issues to which the model has already been applied are indicated by asterisks in the table.

The different types of conclusions we drew concerning REM's policy applicability can be illustrated with several examples from the table. We felt that the model was well-designed to analyze the effects of changes in cost factors such as fuel prices or changes in the regulated rate of return. Examples of such applications exist in the REM literature and are well executed.

On the other hand, the demand component of the model is quite aggregative and does not explicitly analyze end-use appliances. The model, therefore, cannot appropriately be used to analyze the impacts of conservation policies such as end-use efficiency standards. We felt it was worthwhile to point this limitation out to

FIGURE 2

BAUGHMAN-JOSKOW REGIONALIZED ENERGY MODEL

ASSESSMENT PROCEDURES

190
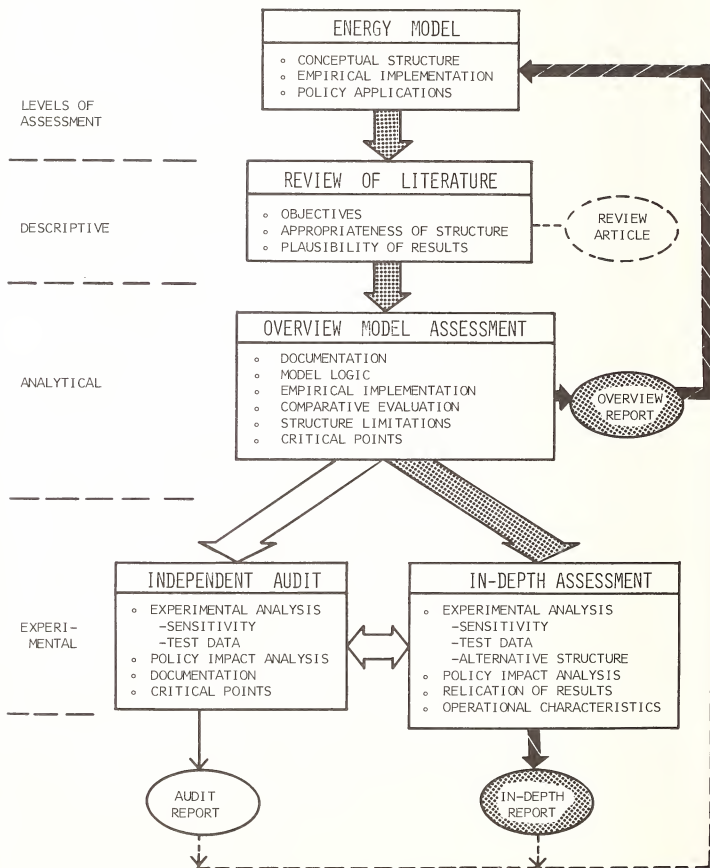
POTENTIAL POLICY APPLICATIONS TO BE CONSIDERED IN ASSESSING
THE BAUGHMAN-JOSKOW REGIONALIZED ELECTRICITY MODEL

1. Changes in factors affecting electricity demand growth paths

   - economic/demographic trends
   - conservation policies

2. Load management

   - peak load pricing*
   - cogeneration
   - seasonal pricing

3. Impacts of changes in cost factors

   - capital costs for new plants
   - fuel prices*
   - wage rates
   - taxes (possibly a Btu tax)

4. Changes in resource supply conditions

   - resource constraints*
   - increasing cost supply schedules

5. Costs of financing*

6. Industry responses to capital "shortage"

   - state financing*
   - less capital-intensive technologies
   - reduce growth
   - reduction in plant reserve margin

7. Regulatory policies

   - regulated rate of return*
   - inclusion of work in progress in rate base*
   - exclusion of noneconomic plants from rate base*
   - regulatory lag

8. Alternative lead times for capacity expansion*

9. Environmental constraints

   - siting restrictions
   - capital equipment requirements*
   - increased operating costs*

10. Technology assessment

   - advanced generation technologies: centralized and distributed
                                       conventional and nonconventional
                                       cogeneration
                                       fuel conversion
   - nuclear: non-LWR, breeder, etc.
   - storage
   - T and D

Applications for which examples already exist in the REM documentation.

potential users even though there are no instances in which the
model has been applied to this type of issue.

To take a third and final example, we had some severe res-
ervations concerning the model's applicability to the analysis
of peak load pricing, even though this was an area in which the
model had in fact been applied. The problem is that policies
such as peak load pricing are essentially designed to change the
shape of the load duration curve but, in REM, the shape of the
load duration curve is exogenously specified. Also, the shape
of the load duration curve is independent of changes in demand.
Even if peak load pricing were to produce drastic changes in
the composition of demand, this would cause no change in the
shape of the load duration curve. Thus, much of the analysis
of the impact of peak load pricing policies has to be performed
outside of REM and then fed into the model as changes in exoge-
nous data inputs. For this reason, we felt that REM was not
directly applicable to the analysis of peak load pricing policies.

Turning again to the assessment process outlined in Figure 2,
after completing the overview assessment we proceeded directly to
an in-depth assessment involving "hands-on" experimental analysis.
The experiments included such things as sensitivity analysis based
on changes in key parameters and data inputs; changes in model
structure; and policy impact analysis.

After completing the experimental analysis, we produced a
final report that included the results of both the overview and
in-depth assessment. This report is due to be released soon by
EPRI. From our experience, we concluded that it would have been
a mistake to have issued a report at the conclusion of the over-
view assessment. A report at that stage would almost surely have
left a confusing, and possibly misleading, impression in the minds
of many readers. When we did the quantitative in-depth analysis,
we found that some of the contention points identified in the
overview assessment were not really critical in terms of their
overall impact on the model's behavior. It is our feeling that,
while an overview report is a useful internal document, a pub-
licly available assessment report should be based on some sort
of experimental data derived either through in-depth analysis or
through an independent audit.

3.   Independent Audit: The Wharton Annual Energy Model

The process of assessing the Wharton Energy Model was, in
many ways, similar to the "audit" process by which an accounting
firm examines the books and annual reports of a corporation. This
type of evaluation is carried out by an independent, outside party
but requires the active involvement and cooperation of the entity
being assessed. The Wharton model was, at the time of the audit,
in the process of development. Documentation was sparse and fre-
quently out of date, with changes being made on a daily basis.
Thus, the audit was like trying to hit a moving, and sometimes
dimly perceived, target. A criticism that was valid yesterday
might be off the mark today. Although attempting to assess a

model which is changing so rapidly can be very frustrating, it is at this stage of model development that an assessment is likely to have some of its greatest payoffs. The audit of the Wharton model provides a valuable prototype for one of the important functions of the model assessment process.

The assessment procedures and their relation to the general assessment framework are outlined in Figure 3, with the shaded portions indicating the steps included in the Wharton audit. As with any assessment, the first step in the audit was to examine all of the available documentation. The verbal descriptions of the Wharton model tended to be quite general, as would be expected for any model in its early stages, but the documentation of the mathematical relationships used in the model was quite good. The only empirical verification that was available was in the form of the statistical measures reported for individual relationships which had been derived through regression analysis. There had been no opportunity to attempt historical replication, sensitivity analysis, or similar tests of the model's properties. There was virtually no documentation of the computer programs used to implement the model since these programs were still being written.

With such sparse documentation, it was decided that an overview assessment was impractical. Instead the assessment proceeded directly to the stage of an independent audit. The audit procedures were conditioned by the two following observations:

- simulation experiments were needed to test the model's properties; and,

- the model developers were the only people capable of operating the model to carry out those experiments.

In addition to the policy simulations that had already been completed by the Wharton staff in support of an EPRI/Stanford Energy Modeling Forum study, about a dozen simulation experiments were conducted specifically for the purpose of generating information for the model audit. These experiments were defined by the assessment staff in consultation with the model developers and were then implemented by the Wharton staff. An important aspect of the audit procedure was that I was present when the Wharton model was run, and the precise implementation procedures were explained to me. "Looking over their shoulders" in this fashion is essential to accurate interpretation of the results produced by the experiments. The outcome of a policy experiment is frequently determined as much by the way in which the policy is introduced into the model as by the way in which the model responds. Furthermore, a great deal of information was gained from discussions with the Wharton staff concerning the ways in which the experiments might appropriately be conducted.

FIGURE 3

WHARTON ANNUAL ENERGY MODEL
ASSESSMENT PROCEDURES
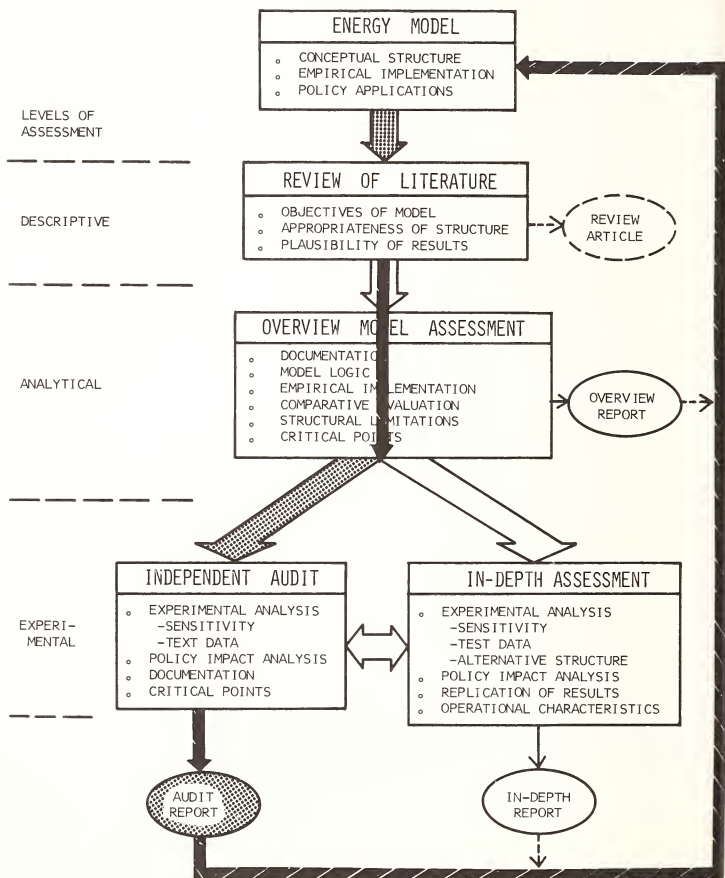
The audit report was completed within two months. Since it was based on a preliminary version of the model, it was distributed only to the modelers and the model sponsor (EPRI). These procedures clearly demonstrate that the audit approach is sufficiently flexible and can be completed rapidly enough to incorporate its findings effectively in the model development process.

4. Concluding Observations

As a summary of the experience gained from the two assessments just discussed, I would like to outline an approach that seems generally appropriate for energy model assessment. An assessment must, of course, begin with a thorough review of the available literature dealing with the model in question. Next, it would proceed to an overview assessment that would use the technical documentation to carry out detailed analysis of the model's logic and implementation procedures. In most instances, the computer code would be a key element of the documentation used in the analysis. A major output from the overview assessment should be the identification of the model's critical points.

I would suggest that most assessments should next proceed to an independent audit and it is only after completion of the audit that a decision should be made about whether to go on to in-depth assessment. Because in-depth analysis involves direct hands-on operation of the model, it is substantially more time consuming and costly than an audit. An audit can function as a relatively inexpensive screening device and in some instances it will turn out that an in-depth assessment is not really required. And, in those cases where further in-depth analysis is called for, the audit will have sharpened the assessors thinking and will provide a sound basis for more intensive evaluation.

Generally a report would be issued upon completion of each stage of the assessment process (overview, audit, and in-depth). However, the experience to date indicates that an overview report, because it raises more questions than it answers, is best treated as an internal document. It facilitates communication among model builders, assessors, and sponsors but can be misinterpreted by readers not directly involved in the process. An audit report would generally be kept as an internal working document for a model still in the development phase but could be made publicly available for more mature models. Since in-depth assessment would only be justified for models that are being used in policy applications, a report coming out of an in-depth assessment should certainly be publicly available.

As indicated by the dashed arrows in Figure 1 leading from the assessment reports back to the model itself, the assessment findings should be a part of the model development process. Indeed, this kind of feedback leading to improved policy models is likely to be one of the most significant contributions of the assessment activity. To facilitate this, I think it is important

that the modeler be involved in the assessment process from the outset. Also, the assessors have an obligation not merely to criticize the model but to suggest ways in which it can be improved. At the same time, the assessors should not themselves implement new "improved" versions of the models. By doing so, they would become competitive model builders and would not be able to function as objective model assessors.

Finally, I would like to suggest that, because model development is a continuous process, model assessment should not be terminated with the completion of an audit or in-depth report. When a model is improved, perhaps in response to the initial assessment, an updating of the assessment should be undertaken. This update would probably take the form of an audit and, because of the information base and expertise established during the initial assessment, the incremental cost would generally be very low. The credibility and reliability of energy policy models could, in my opinion, be significantly enhanced by thus including rigorous, objective assessment as a standard element of the modeling process.

REFLECTIONS ON THE MODEL ASSESSMENT PROCESS:
A MODELER'S PERSPECTIVE

Martin L. Baughman
Center for Energy Studies
The University of Texas at Austin

At the time I was approached about making the Regionalized Electricity Model (REM) available for assessment, we agreed that such an activity was desirable--indeed essential--for the advancement of the energy modeling profession. At the same time, I felt that the Regionalized Electricity Model was developed to a point where scrutiny by a third party would prove beneficial to further development of the model; it would, as well, make the model transparent, and thus useful, to potential users. And though it was not without some trepidation that I offered the model for assessment, I felt at the time that this particular model would be a good trial for the assessment laboratory.

The modelers still disagree in some areas with the MIT Group's presentation of the model; these areas are delineated later in the paper. Before setting forth these details, however, I would like to comment generally on the issues raised in this first independent assessment. MIT has labeled and listed these as follows:

- (i) the extent to which the models being assessed should be compared to similar models;
- (ii) formalization of the relationships among the modelers, the assessors, and the sponsors;
- (iii) approaches to independent assessment; and
- (iv) the nature and extent of in-depth independent assessment.

Individual vs. Comparative Model Assessment

What _really_ is the distinction between an individual assessment and a comparative assessment? The distinction between these two modes of assessment is not made very clear in the MIT work. The MIT assessment of the Regionalized Electricity Model states on page 1-10 of the draft report: "In the present case it has not been possible to provide a comparative assessment between the Baughman-Joskow model and potential competitors." Although not explicit, this statement implies that what they attempted was an individual assessment, not a comparative assessment. But here a problem exists. The MIT Group states on page 3-56, in the section entitled Electricity Generation: Model Assessment:

> It is worth noting that the generation simulators used by electric utilities are considerably more sophisticated than the electricity generation

> model in REM. The utility models commonly employ
> probabilistic simulation, incorporate many more
> types of generating plants, and take into account
> seasonal factors. The use of an annual load
> duration curve in REM, although a reasonable
> simplifying assumption for some purposes, undoubt-
> edly restricts the applicability of the REM results.

The passage does not state for what applications the model is
restricted as a result of the simplifying assumptions. The passage
clearly states that REM cannot be used for some of the purposes of
the more detailed utility models. I do not argue with the conclusion
at all, but what I don't understand is why the statements exist in an
individual model assessment.

Another example of the same point is the following: On page 3-12
of the MIT report, the first sentence of the section on the demand
submodel entitled <u>Overview Evaluation</u> states:

> The REM demand submodel generally represents the
> state-of-the-art in overall energy demand modeling
> at the time it was constructed.

This section of the report goes on:

> REM does, however, differ in some details from
> other efforts. It lacks, for example, the richness
> of policy variables and technological specificity
> found in other interfuel substitution models.

The report then proceeds to describe capabilities of other interfuel
substitution models and contrasts these models with REM. Again, as
in the case of the assessment of generation simulation, the MIT Group
has used other models as a reference for comparison.

The report then goes on to compare and contrast the financial/reg-
ulatory submodel in REM with the Fishbein model. Again, the MIT Group
has used the standard of another model to make comparisons and to
facilitate the conduct of its assessment.

Now, what is the point of all this? The MIT Group claims they
did not do a comparative model assessment. Yet, here are three obvious
examples where other models were used as a reference for comparison. I
argued from the inception of this assessment activity that a compara-
tive model assessment was the only realistic perspective that could be
adopted for assessment of large-scale energy policy models. Since the
MIT Group has conducted its assessment, I feel even more strongly
about this point. In fact I have concluded that a comparative assess-
ment is really what the MIT Group conducted. Whereas the MIT Group
might purport to have used reality as a frame of reference, in fact it

is the understanding of reality as expressed in models--the sum total of more detailed, more specific, and other large-scale models--which provides the frame of reference for their assessment. To form an informed judgment of a model's structural and empirical appropriateness in describing reality is simply to compare the model with other models. Thus, in fact a comparative assessment was really conducted.

I continue to believe that the real world, as an explicit frame of reference for conducting the model assessment, is an unrealizable and unarticulated standard. I think the MIT Group, at a minimum, should confess to what they are really doing. Further, I would suggest in the future that two standards of comparison should be used. The first might be the unarticulated state of the art implicitly used by the MIT Group. The second is a set of explicit analytical capabilities, perhaps those used by the Department of Energy in its official policy analysis activity. After all, the Department of Energy is the public agency where energy policy in its broadest definition is analyzed.

Relationships among Modelers and Assessors

As a result of weaknesses in organizational relations among the model assessment group, the modelers, and the model assessment sponsor, the second in-depth assessment activity is being conducted in a slightly different way from the assessment of the Regionalized Electricity Model. It became apparent in the first assessment that the demands placed upon the modeler are not negligible. In the second model assessment (and, I recommend, for all future assessments) the modeler is contractually included in the assessment to facilitate interaction and to provide recompense for the time involved.

A very serious question still remains, however, about what piece of computer code really represents "the model." This issue has been placed under the topic of appropriate interaction between the assessment group and the modeler in this report. One position is that the model being assessed is the code that is initially handed over to the model assessors. Future interaction between the modeler and the assessor then is merely on questions of understanding. The MIT Group chose to use the initial computer code in this assessment. I find this arrangement completely satisfactory for a model that is well fixed in its structural and empirical content, but I am hesitant to endorse this conclusion in the case of a model which is a continuing research tool, such as the Regionalized Electricity Model.

A case in point is the following: During the period that MIT was conducting the assessment of the Regionalized Electricity Model, the modelers discovered a coding error in one of the statements relating to the calculations of the economics of alternative capacity types in the capacity expansion portion of the model. This "bug" in the model was corrected in our version, and we modelers brought it to the attention

of the assessment group in our response to their draft overview report prepared in the summer of 1977. Yet, the MIT Group chose not to correct this error in the code in their version of the model before the in-depth assessment. There are sensitivity analyses reported in the in-depth assessment in which the obvious explanation for the behavior encountered is the computer bug that was brought to the assessors' attention. The behavior would have been different had the corrected version of the model been used.

Should the MIT Group have done the sensitivity analyses with the corrections in the computer code? They were informed of the problem by the modelers, not vice versa. Perhaps the assessment of a model should await the completion of the research of a model's development (to the satisfaction of the developers) before it is released for assessment. Or, perhaps there needs to be a follow-up assessment, say six months after the initial assessment, describing changes and corrections implemented in the computer code to verify and document model corrections and improvements.

Finally, an equally serious question is what version of the assessment report really represents the final report. We authors were asked to respond to an overview assessment reported in the summer of 1977. This first assessment report was seriously flawed in a number of ways, and we prepared and sent to the assessors detailed comments and reactions. Shortly thereafter, various versions of the final report were sent to me with requests for review. The first version incorporated the reports of the in-depth assessment in appendices. A few reactions to this material were delivered by means of telephone conversations. We modelers were then informed that the report was being revised. Thus, we made no effort to review this "first draft" of the final report in more detail.

Then, the modelers were presented with another draft version of the report (specifically Chapters 2 and 3) transmitted on May 15, 1978, prior to a project review to be held in Palo Alto on May 25-26, 1978. No written comments were delivered to MIT on this draft; however, I made a verbal presentation of reactions at the review meeting. Requests by MIT for us to prepare a written chapter of responses to the final report were then unfilled until a complete copy of the final report, including the Executive Summary, was available for review. This report was transmitted to me on December 12, 1978. The letter of transmittal indicated that no additional changes to this version of the report were planned and requested written comments on the report to be included as Chapter 4 in the final publication.

I prepared my overview and presented it verbally at the Workshop on Energy Model Assessment held at Gaithersburg, Maryland, on January 10-11, 1979. MIT participated in the workshop. After the presentations Rich Richels, the EPRI project administrator, requested that MIT make further revisions to the report, especially Chapter 1. The copy revised

in response to the Richels request, is, to my knowledge, the copy before you. It was transmitted to me on February 7, 1979, after my reactions were presented verbally with Richels and representatives of the MIT assessment group present at the January 10-11, 1979, workshop.

When I reviewed the final transmittal after the Gaithersburg workshop, several changes in the report became apparent. The Executive Summary and Chapter 1 were changed significantly. Section 3.5 was changed significantly. References in the text to other models contained in the December 12, 1978, version of the report were deleted. The reporting process has to be more structured in the future if the modeler is to participate in fair exchange.

Approaches to Assessment

The third issue raised by the MIT Group comes under the heading Approaches to Assessment. Initially, assessment of policy models was conceived as two alternative approaches: 1) overview assessment and 2) in-depth assessment. The fundamental distinction between the two is whether or not the assessment group actually operated the model and controlled the associated data base.

At the presentation of the draft final report on the model assessment by MIT at Palo Alto last year, I expressed my reservations about an overview, and only an overview, assessment. My experience was that at the stage of the activity when only the overview assessment was completed (i.e., when only documentation of the model had been reviewed) there were a large number of misunderstandings of the model behavior and the model representation that would have been particularly detrimental to the reputation of the model and the modelers, if the assessment activity had been terminated at that point. So many inconsistencies existed between the assessors' understanding of the model structure and behavior and the modelers' understanding of the model structure and behavior at that stage that the modelers could not support a proposal to undertake only an overview assessment.

The original conceptions of assessment needed to be and have been altered as a result of this first experience. Does it mean that an assessment of a model must be a full-fledged, complete in-depth assessment to be worth the effort? Is the independent audit plus overview a reasonable compromise? So little of MIT's final report is devoted to the independent model audit concept that it is difficult to form an informed judgment.

Nature of In-Depth Assessment

As the MIT Group points out, one way to conduct an in-depth assessment may be to exercise the capability to operate and execute the model experiments to replicate previously published results. At the other

extreme, in-depth assessment might be interpreted to mean complete replication of model data, parameter estimates, computer codes, and the results of published applications. The original plan for the in-depth assessment of REM called for the following:

1. Checking independent data used in the model back to primary sources
2. Replicating estimated parameters
3. Estimating new structural relations where technical results are questionable, including them in the model, and performing sensitivity analyses to determine if published analytical results might be compromised
4. Verifying computer procedures and codes through analysis and recoding
5. Replicating unpublished analytical results.

However, in the case of REM, the MIT Group states (p. 1-17):

> We modified our original, rather extreme concept of indepth assessment to focus upon verification of computer code and sensitivity analysis of the key parameters and independent data identified during the overview analysis.

The measure of success of the in-depth assessment is, in my opinion, however, somewhat inconsistent. In practice, then, the in-depth assessment really set forth only very limited objectives, and the label "in-depth" is misleading.

Sensitivity analysis cannot substitute for discussion of model validity. In the overview assessment of the supply portion of the model, the MIT Group criticized the supply submodel as possessing several biases and not really representing a good description of electricity production and capacity planning practices. But the sensitivity analyses did not illuminate this point. The question of how the industry would behave under the same controlled conditions imposed in the sensitivity analyses was not addressed. The behavior of the model was illuminated, but the validity of the model was not.

I think the assessment report offers much insight and an informed point of view on the Regionalized Electricity Model, but at the same time the model assessment is not entirely above reproach. This first in-depth model assessment as manifest in this report sets a high standard for future assessments.

As a result of the overview assessment of the model that was completed in mid-1977, several changes were implemented in the Regionalized Electricity Model. These included a correction of all the computer "bugs" relayed to us as a result of the MIT effort to reprogram the model and changes in the way the model reported capital shortages.

Other changes are being made in response to recommendations of the in-depth assessment, but, as might be expected, making these changes involves more time and effort than those implemented in response to the overview assessment. As a result of the assessment activity, I think that the Regionalized Electricity Model is now more transparent than it was before the assessment took place. Its behavior is better understood and its limitations are more widely known.

DISCUSSION

Dr. Greenberg (DOE): This description of third-party assessment raises what I think is an issue that I would like to pose for the record. The issue is what I think is an intrinsic limitation of third-party assessment as a technique to try and get at a measurement of the quality or the usefulness of a model.

The limitation comes from a separation that I think is inappropriate. One of the early applications of the Baughman-Joskow Model was a study of impacts of a nuclear moratorium. And I think that if somebody other than Marty had used the model to conduct that analysis, the quality would be different. So I raise the issue that you can't separate the people when you talk about a model.

I think a model as we use it in our field, is to transform information from one form to another for the purpose of enlightenment. In that regard, I think a model includes the inanimate portion, in the form of a computer code, which is what third-party assessment seems to be assessing. There are two other things that happen. The way the model gets used is for somebody to ask a question; then, a person translates the English into modeleese, which is the input specifications for scenarios designed to help answer the question. The inanimate portion then translates the input modeleese into the output modeleese; then, a person interprets the answer and translates the modeleese into English.

Usually there are iterations where, after you get by the laugh tests, and things get a little bit subtle and the model does something surprising, you have to go in and find out why the model did what it did. Then you either find a faulty component, repair it and repeat, or you revise your intuition and have what you think is a believable story in English. That is usually what takes place.

Therefore if you buy that, I have got a bridge for sale, but, also if you buy that, I have got what I think is a major factor in determining the quality of results and the usefulness of the whole model. That is, you have to have intimate familiarity with the model to use it properly.

Somebody using it as a black box can never be of the same quality as somebody that is intimately familiar with the model, which is why I think modelers and analysts are inseparable. The analysis should be conducted by the people intimately familiar with the model so the interpretation is made correctly, and the translation process is done with the highest possible quality. It is this inseparability of the people that use the model by being intimately familiar with its innards, that I think offers an intrinsic limitation on what third-party assessment of the inanimate portion can achieve.

Dr. Kresge: I think I would at least like to respond to it in passing, because it seems to me that the conclusion you draw from that, especially since in my other incarnations I am a modeler, it is very encouraging because it says, hey, don't assess me because I am inseparable from my model, and especially don't assess my model, because if you do you are doing it an injustice.

If we really believe what you say, then it seems to me that we either have to give up trying to assess anything, or you have to suggest an alternative procedure. Then on top of that, I don't see how that leads to a way of judging between competing models done by competent people that produce different results.

Dr. Greenberg: I just want to say one comment in response to that. I said I think it offers a limitation. I didn't say I think it renders it valueless.

Dr. Richels: Okay, I would like to add one more point. I have thought a lot about the problem of separating models from modelers, and I think in the last couple of years I have probably turned around 180 degrees. I question the feasibility of separating the model from the modeler for policy analysis. I don't question though the desirability of separating the model from the modeler for model assessment. And I think that the MIT group is intimately familiar with the Baughman-Joskow Model now, and one reason why we want to build in the participation of the modeler, Marty Baughman, in the assessment explicitly is to prevent us from going off in the wrong direction when indeed that happens. And, therefore, the differences are differences of opinion of modelers of analysis to particular aspects of the model and not whether the model is behaving in a particular manner.

Dr. Baughman: I agree with the remarks that Rich Richels just gave on this topic. I think that the question is a legitimate question. For example, in the in-depth assessment that MIT conducted, had I been present there were explanations for results--maybe not very good explanations, but explanations nonetheless--for why, when you jiggle those inputs, those outputs came out.

And so, to that extent, I think you are right that you can't separate the modeler as an analyst from the model. I think Rich gave the appropriate response. I think, as a result of this first activity, that it has been pretty well concluded that the modeler has to be included as part of this process, and so that is recognized in future activity.

Dr. Sweeney: In thinking about model assessment, I have decided that one of the most difficult things to catch is the implicit assumption, the difference in the world view between different modelers, or the world view incorporated in the model. Has the process that the MIT energy lab has gone through helped sort out some of these implicit assumptions? I will give an example from an energy model forum study that we went through. This is something that came up in a comparative study of a number of models.

The coal and transition study noted differences between geographic patterns among several of the models. It was traced back to the fact that an implicit assumption in one of the models was that there was a monopoly in the supply of coal.

I mean that was sort of implicitly in there and the modeler didn't understand that that was in the model. It was subsequently changed as part of the process.

Is there anything about the MIT energy assessment lab process that helps you get at that, or is it more in getting at sort of the guts of the model's explicit assumptions, the explicit coding issues?

Dr. Kresge: I think that there have been a couple of comments on that already. Marty made most of them, I believe. He was pointing out, that especially in the overview portion of the assessment, as opposed to the in-depth, comparative analysis inevitably is in there; you are using people who have substantial expertise in the field.

They are aware of other studies and those are brought into play. Rich made the point, that I think we all agree, that it would be very nice if the resources were available to do comparative analysis of key models within any given area. If we could afford to do comparative in-depth analysis of several coal models, that would be a very, very desirable thing, and hopefully that is something we are working toward.

Let me emphasize another point that is sort of peripherally related to yours, and Marty brought it up, and I would like to stress it. From the outset, we recognized that there were going to be points, like these differences in implicit assumptions, that would cause the assessors and the modelers to end up with irreconcilable differences. We would say that something was a limitation of the model and that the modeler would say was right. To deal with that--at least at one level--we had in the initial experiment, and we have on a continuing basis, a very firm rule that in the final report there is a chapter there where the modeler has a chance to respond to the assessment.

The report cannot be on an assessment basis only. There must be a chapter in there where the modeler can respond and say, look they have this world view. I have this other world view. We agree to disagree.

Apart from doing more than one model at a time, the only way I see the comparative assessment coming in is through the other studies that we are aware of. And, of course, it would be nice if there were more coordination with things like the forum.

Dr. Sweeney: Thank you. May I make a quick point to sum up what is at issue? When I say comparative model studies, I don't mean look at one study and say, well, that is pretty good because they include that, and this one doesn't include that. I mean standardizing the input to see how each of them behaves to a standardized set of inputs. That is what, I think, brought out the differences that we found.

Dr. Kresge: That is certainly what you would do, say, if you are doing side by side in-depth analysis. You would certainly do it by standardized experiments, so that that would automatically be in there.

Dr. Richels: I would like to add one point to what Dave said. I think he has brought up a very interesting feature. That is that by the time we get the final report at EPRI and we have had the final report there for a while and have been getting Marty's input all along, we do not take the opinion of the assessors as the gospel that factors into our decision when we are looking at a particular application of the model. In no way do we try to reach any kind of consensus between Marty Baughman and the assessment lab, nor do we encourage any consensus except over factual disputes. The kinds of inputs that arise through their disagreement is, in my opinion, the most valuable part of the process.

Dr. Manove (Boston University): I have been somewhat associated with MIT in their assessment. I have been sort of very disturbed by a few things I have heard and I just want to comment on them and also by something that I have done, which is the assessment project.

I have found myself in assessing the ICF model moving more and more toward verification and less and less toward validation, as you have defined it here. I think the reason that I have found myself doing this and my colleagues doing this is, this nonsense that we cannot know about the real world, and that all we can do is sort of give up and check each formula and see if it is consistent and check the data to see if anybody has made a mistake in adding up some numbers or compare this model with another one.

We can know about the real world and, if not, I think we should all quit and go home. What we have been saying here reminds me of people that say that our winning evolution is a theory, not a fact or people that might say that the assumption that the world was round, is somebody's model that explains the things we see. Sure, the round world is a model. Pre-dicting an eclipse in 1992 is a model, but those are models that we know so well and we believe so closely that we call it a fact. We say that this is really true about the real world.

There are things that we do know about the world out there and what we ought to do in validating these models is sit down and figure out what do we really know. What do we really believe about the world. And in those few areas where we do know something, these models jive with what we know.

Now it will take some doing to figure out what we do know about the world but it can be done. One of the things that I think we do know, at least we believe strongly, is that the world is fairly continuous. That if you change a price by half of one percent, the quantities do not change by a factor of forty. That is why we do these activity analyses. That is one reason why you do it. You move something a little and if the model goes crazy, you say, hey, that must be wrong. We do know something about the world; we know this continuity thing.

There are things that we can test by ordinary empirical tests. For example, in the ICF model, they assume that the solution is going to be cost minimization for the whole country. Well, we can find out whether our costs are being minimized right now, by utility. We can do that kind of testing; we cannot find out for certain, but we can try to say something about the world.

I really think that there are some things that we know; there are some things that we know with a high degree of confidence and we ought to think about what we know and then we ought to really try to test validity, not just to compare things.

I think we ought to not be satisfied, either, with just testing validity of little pieces of the model; we have to test the validity of the whole model. If you test out each little piece as Hoff was describing, and sure you find out about transportation, but maybe the whole damn thing is wrong because your whole conception--the whole way you put it together-- is wrong. You might be subject to some catastrophic error in your whole model. So, you have to do more than find out about a lot of little pieces; you have to ask big questions.

I think my experience with MIT has been not to ask the big questions. We have all been trying to push ourselves to ask big questions that we really believe this thing and does it really jive with what we know about the world. I hope that this kind of a "Gee, we cannot know anything about the world, anyway" attitude goes away.

That is my little speech.

    Dr. Marcuse (Brookhaven): I assume that the reasons we do validations is really an attempt to improve models. Out of the validation process we hope to get improvements in the models, not weaken them.

I guess I would define improvement probably that the model supplied better information. Then the question is, what is better? I suppose the answer to better would be, is there some way it now gets used in the decision process in a way that it has an impact on making the decision. Hopefully, a better one.

The question I have is in what way have any of the models that they have assessed up to now been improved as a result of the validation?

Dr. Baughman: I want to respond, I guess, in a couple of ways to the comments and questions that you have made.

First of all, it strikes me that a model, I still agree, cannot be proven whether it is valid or invalid. A valid model, I think by definition, however, is one that is devoid of contention points, another word that has been used here today. And that if the model has no contention points, then it is pretty well accepted that the behavior of the model conforms to how everybody believes the real world behaves.

I still believe that whether or not it is valid still cannot be answered. F equals ma looked very good for a long, long time until something better came along, but it was a perfectly valid model for most of the applications to which it was put before relativity came along.

In terms of response to the question, how has the model been improved, there were several errors in programming that were brought to our attention as a result of efforts to reprogram the model. These have been corrected.

I think that if I may paraphrase and briefly point out what I think MIT's report said, that if you look at the three basic components of the regionalized electricity model, in terms of their relative quality, probably the most interesting part of the model was the financial regulatory component; the part that was kind of so-so was demand; but, probably the weakest part in a comparative sense was the supply sub-model. That is useful input. I think they probably knew that, but many times that information goes subliminal and you do not explicitly discuss these things, especially with those who might be potential users of the model.

As a result of that, efforts have been made to reprogram much of the supply portion of the model. A lot of that is completed; some of it is still to be completed before some additional applications of the model are being made. We have and are responding to suggestions that were made in the assessment.

Dr. Richels: I think that Marty has responded to one side of the coin and that we are looking for feedback from the modelers and, hopefully, an improved model or more useful model. The other side of the assessment process, though, is to aid the model users in the intelligent use of the model. I can think of several instances at EPRI over the last several months where we have had discussion involving the use of the Baughman-Joskow model. I can think of a few instances where the model assessment was quite helpful in assisting us in determining how we were going to use the model.

Dr. Greenberger (Johns Hopkins University): I just wanted to clarify this term contention point. At least, as it was originally intended.

It is not something that you want to eliminate in models, it is really something you look for in models. In particular, you look for points in models where there is an ambiguity, a possibility for different assumptions which corresponds to an area of disagreement in the policy field that you are studying so that the model gives you some way of exploring the different assumptions possible in tracing back to basic differences in points of view.

Now, if you are studying a particular policy area with a particular model, you would like your model to reflect the actual contention points. If the model, instead, essentially resolves those contention points; in other words, makes an assumption which fixes the way the dispute in an actual case is represented, then at least for exploring that difference of opinion it is useless.

An additional concept that was used along with contention point, although I have not heard it referred to today, is the notion of a contention point being critical as opposed to noncritical. What that was meant to signify was if changing the assumption in the model alters the policy consequences, in other words leads to one policy action versus another, then that contention point is critical in the model. The contention point is not critical if it really does not matter which of two possible assumptions you make.

Just to give you an example, in the model which was being assessed for which the concepts were originally devised, a contention point was whether or not the oil and gas industry was competitive. And that turned out to be a very critical contention point, although the assumption in the model was that it was competitive. Essentially, the contention had been resolved, so, therefore, unless you had a model which could explore a competitive industry versus a non-competitive industry, you could not really get at the root of the disagreement. In this case, the policy area was deregulation of natural gas prices.

Dr. Stauffer (ICF): Marty, do you think, on balance, the process was a net benefit: a) to you; b) to the rest of the world? Then, was it worth the cost? And finally, could it have been done more efficiently?

Dr. Baughman: Well, I am not sure that the book is closed at the present time, but my perception is at this point that, yes, it was a benefit to me in terms of organizing research priorities and even suggesting some new research areas. Since that happens to be one of the things that I pursue in my professional endeavors, that is worthwhile.

To the other questions, I would have to let others judge. I feel that from what I have seen and the assessment report the MIT group has put together, that much of the mystery of the model, if there was any before, had to be removed as a result of this activity. I would like to see the activity continued.

THE TEXAS NATIONAL ENERGY MODELING PROJECT:
AN EVALUATION OF EIA'S MIDRANGE ENERGY FORECASTING SYSTEM

Milton L. Holloway[1]

INTRODUCTION

Large-scale models are a product of our times and seem to be growing in
the importance of their use in government policy work. Policymakers are using
large-scale models and really have no alternative due to the complexity of
potential impacts from policy decisions but at the same time they are sceptical
of the reliability of forecasts and calculations from models. There is also
great acceptance by the populace at large of results coming from computer
analyses which in their minds seem to represent the epitome of technological
solutions to problems. Analyses are somehow seen as more believable if they
are based on computer technology.

The major task that lies before us is to improve the usefulness of models
and the judgments of professionals involved in public policy analysis. The
central issue is the procedures by which the reliability of large-scale models
can be established and made transparent--to distinguish between the influence
of professional, subjective judgments and the influence of objective informa-
tion that is reproducible by others. The Texas National Energy Modeling Proj-
ect (TNEMP) has made some contribution to the goal of increased model
credibility by transferring and operating the Midrange Energy Forecasting System.

PROJECT PURPOSE

The first purpose of the Texas National Energy Modeling Project (TNEMP)
is to provide an independent evaluation of the Energy Information Administra-
tion's Midrange Energy Forecasting System (formerly known as PIES). The
evaluation will provide guidance to users of MREFS concerning the level of
confidence one may have in the results of the models for government energy
policy analysis purposes. The evaluation is critical, in the best sense of
the term as used in scientific work, but also makes helpful suggestions for
improvement in the model structure and in procedures used by the Energy Informa-
tion Administration (EIA) for increasing model credibility.

The second purpose of TNEMP is to provide recommendations to the Texas
Energy Advisory Council concerning the maintenance of a national modeling
system by the Council for purposes of evaluating Texas impacts within a con-
sistent national modeling framework. As a result of the exercise, we have
first-hand experience with MREFS, as well as DRI's Macro and Energy models. We
are able to compare these models, their structure and results with various
Texas models at the Texas Energy Advisory Council, the University of Houston
and the University of Texas for purposes of assessing their relative usefulness

---

[1]   Dr. Holloway is Executive Director, Texas Energy Advisory Council, Austin,
      Texas, who **serves as** Project Director for the Texas National Energy Modeling
      Project.

for energy policy analyses and possible joint use or integration. TNEMP participants have experience in still other modeling efforts, as well as experience in model evaluation, institutional arrangements for housing models and procedures for making the best use of resources to achieve successful model development, use and credibility. TNEMP is well suited for the second purpose.

The evaluation of Texas impacts of national energy policy decisions must necessarily be done within a consistent national framework since a major portion of the nation's energy production and processing as well as corporate management of these for the nation as a whole is in Texas. A significant fraction of the nation's energy is also consumed in Texas, especially natural gas. Texas based energy analyses must, therefore, be national to properly reflect Texas' interests and its role in the nation's energy future. In order to be of use in the national energy policy process Texas based analyses must be centered in a credible national modeling and analysis framework.

A third purpose of TNEMP is to raise the level of attention concerning the current uses, practices, potential as well as the current abuses, and potential dangers of modeling for purposes of developing public policy and clarifying important issues to the citizens of a democratic society. The increased reliance of policymakers, high level advisors, and the voting public on expert opinion, based in part on large-scale data bases and models, in an era of complex problems and policy prescriptions, requires that policymakers and high level experts gain a better understanding of current modeling reliability and practices. TNEMP is intended to help achieve this goal.

Two additional factors are important side effects of the TNEMP exercise. First, we have advanced by some degree the art of model evaluation by achieving, among other things, the transfer and operation of a large and complex model. Hopefully others will benefit from both our successes and our mistakes for model evaluation is far from being a well-developed discipline. Second, we hope that some insight has been gained into what kind of institutional setting is appropriate for third party model evaluation efforts. The current institutional arrangements for the building, operation and maintenance of large models are critically lacking in efforts aimed at validation of relations, assumptions, data verification and model documentation. The quest is for the development of model evaluation institutions and/or incentives to bring about a timely devotion of model developers to validation and verification and to the provision of clarity and workability as model attributes. Third party evaluation can increase the probability of a model being accepted. It can enhance the transfer of science, technology, and statistical data to the policy decision process.

ORGANIZATIONAL STRUCTURE FOR THE STUDY

The organizational structure for the study consisted of four primary groups. First, the Texas Energy Advisory Council Executive Director provided project direction and the staff provided coordination, materials and other support. Second, the National Advisory Board provided advice on procedures, suggestions on methodology, evaluation of the Analysis Team results and general recommendations for a Texas national modeling capability. Third, the Analysis Team provided an evaluation of MREFS, made recommendations for improvements and alternatives, and made specific recommendations for Texas

maintenance of a national modeling capability reflecting the impact on or by Texas. Fourth, the Supporting Institutions provided support by endorsing the objectives of the study, making review and comment, making data and facilities available, and funding the study. Figure 1 illustrates the working relationship of the four groups.

Meetings were held periodically for the purpose of reporting progress of various studies to the National Advisory Board, other members of the Analysis Team, and the Supporting Institutions. This format provided opportunity for refinement of project objectives, definition of evaluation criteria, sharing of reference material, identification of weaknesses in the project study design, interaction of project participants with DOE personnel and overall guidance for the project director and individual principal investigators.
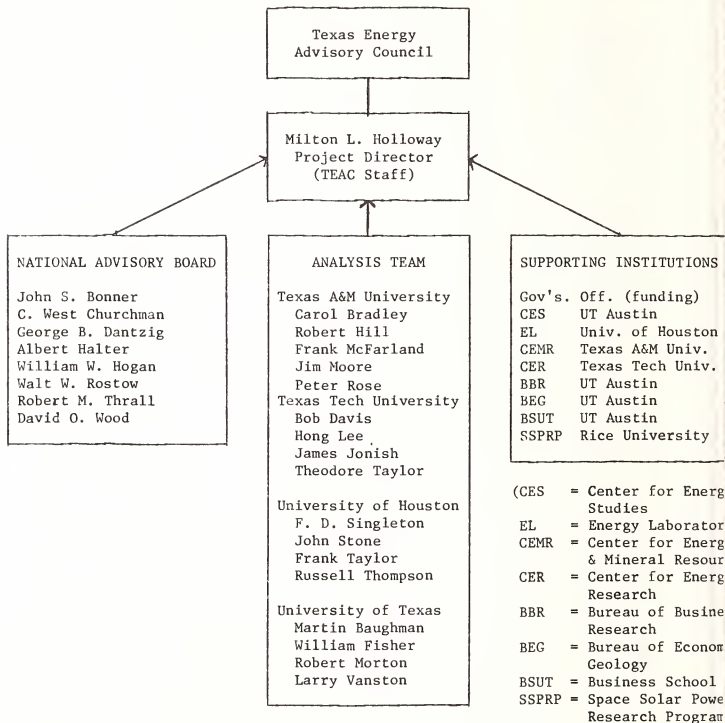
In order to provide credibility of the project objectives and procedures, the National Advisory Board was asked to write an evaluation of the study to be published with the final report and their statement is included at the end of Volume I of the TNEMP report. In order to encourage maximum intergovernmental cooperation, EIA was offered the opportunity to comment on the project with the assurance that the remarks would also be published with the final report. EIA comments are also included at the end of Volume I. To achieve the objective of familiarizing key decision makers with the important findings of the project a meeting is now being arranged between members of the National Advisory Board, officials of EIA and the Department of Energy, and the Lieutenant Governor of Texas, who serves as Chairman of the Texas Energy Advisory Council.

EVALUATION CRITERIA

The reliability of information from the EIA large-scale energy model known as MREFS, currently being used for forecasting, policy evaluation and policy analysis, is our primary concern in this study. The determination of reliability is a difficult task; much has been written recently on the topic but not much is well defined and clarified at this time. We have made certain interpretations concerning the process and measures to be used in assessing reliability. In the usual understanding of reliability in scientific areas one would expect modeling results to be unambiguous, reproducible, and transferable. Such characteristics need to take on special definitions however, since we are dealing with a system involving the behavior of people (the economy with emphasis on the energy sector) and are making applications of models to describe alternative worlds, some of which will never exist for testing purposes. We also need to be specific with respect to the user of the information and in our case this will involve both other modelers and professionals as well as policy decisions makers and the public at large. It is also essential to distinguish uses of the models as between raising issues versus resolving issues. Models used for raising issues may rely more on hypothesis for their formulation whereas models used and designed for resolving issues will necessarily have to be based on accepted theory and/or laws.

MREFS of the Energy Information Administration (EIA) is used primarily for the purpose of resolving energy issues and their use is for policymakers and the public at large. For that reason the evaluation criteria we have selected for TNEMP take on specific definitions.

Figure 1.
Texas National Energy Modeling Project
Organizational Structure


```
                    ┌─────────────────────┐
                    │   Texas Energy      │
                    │  Advisory Council   │
                    └─────────────────────┘
                               │
                    ┌─────────────────────┐
                    │ Milton L. Holloway  │
                 ┌─▶│  Project Director   │◀─┐
                 │  │   (TEAC Staff)      │  │
                 │  └─────────────────────┘  │
                 │           ▲               │
        ┌────────┘           │               └────────┐
```

NATIONAL ADVISORY BOARD	ANALYSIS TEAM	SUPPORTING INSTITUTIONS
John S. Bonner	Texas A&M University	Gov's. Off. (funding)
C. West Churchman	Carol Bradley	CES UT Austin
George B. Dantzig	Robert Hill	EL Univ. of Houston
Albert Halter	Frank McFarland	CEMR Texas A&M Univ.
William W. Hogan	Jim Moore	CER Texas Tech Univ.
Walt W. Rostow	Peter Rose	BBR UT Austin
Robert M. Thrall	Texas Tech University	BEG UT Austin
David O. Wood	Bob Davis	BSUT UT Austin
	Hong Lee	SSPRP Rice University
	James Jonish	
	Theodore Taylor	
		(CES = Center for Energy Studies
	University of Houston	EL = Energy Laborator
	F. D. Singleton	CEMR = Center for Energy & Mineral Resour
	John Stone	
	Frank Taylor	CER = Center for Energy Research
	Russell Thompson	
		BBR = Bureau of Busine Research
	University of Texas	BEG = Bureau of Econom Geology
	Martin Baughman	
	William Fisher	BSUT = Business School
	Robert Morton	SSPRP = Space Solar Powe Research Program
	Larry Vanston	

214

There is currently no consensus of opinion on either the appropriate set of criteria for evaluating models or the definitions of commonly used terms in model assessment literature and discussions. Various discussions of model assessment criteria may be found in recent publications by the MIT Model Assessment Group, Saul Gass, Judith Selvidge, A. N. Halter and others.[2]

Four common themes run through the discussions however, and form the basis for our evaluation criteria in TNEMP. The criteria are workability, clarity, validity (coherence) and verifiability (correspondence).

The four themes commonly discussed in connection with model assessments are interrelated. First, if the model is to serve any useful function other than the satisfaction of the intellectual curiosity of the modeler then it must serve some practical function—identify issues, suggest a resolution of issues or in general it must produce results that relate to the practical problems of the user. We have interpreted this theme as our criteria of workability. In the context of the energy model which we are evaluating we interpreted workability to mean that the model was capable of providing information regarding such things as the inflationary impacts of decontrol policy, the impacts of conservation and decontrol on the import levels of crude oil, the effect of the combination of energy policies on international trade and the value of the dollar, the impacts of deregulation, coal conversion and other such policies on economic growth, and perhaps the regional distribution of economic growth, as a result of national policy. So in the context of our model evaluation effort the model is workable if it raises important policy issues in these areas, points toward the resolution of important issues, or provides explicit information pertinent to an issue.

Second, it is implicit that a model cannot have good workability characteristics if it is not clear. Our interpretation is that clarity must be defined in terms understandable to other modelers (who, for example, may provide a function of advisor or interpreter to laymen and/or policymakers) and in terms understandable to laymen and/or policymakers, who are the users. For our model evaluation effort this means that the model's behavior and its results must be translatable into various supply and demand representations of the economy, understandable to economists, and intuitively comprehensible by laymen who want to know the inflation, income, tax, energy costs, employment and economic growth implications of the policies being analyzed. This understanding is embedded in our use of the criteria of clarity.

Third, if a model is to have workability and clarity, it must be unambiguous in that the model behavior must correspond to the modeler's expectations as per its design. That is, the model must do what the modeler says it will do. It also follows that the model must behave in a manner consistent with the underlying logic or theory upon which it is built. In terms of our model evaluation we should be able to verify that increased prices for oil and

2   See for example, "Evaluation of Complex Models" (Gass, Saul, 1977),
     "A Procedure for the Evaluation of Complex Models" (Gass, Saul, 1977),
     "Strengths and Limitations of Models from a User's Viewpoint: Modeling
     the Modeler's Model" (Halter, A.N., 1977), Texas National Energy Modeling
     Project: Volume I Project Summary (Holloway, Milton L., 1979),
     "Independent Assessment of Energy Policy Models: Two Case Studies"
     (The M.I.T. Model Assessment Group, 1978), "Panel Discussion - Management
     Audit of Quantitative Models" (Selvidge, Judith, 1978).

gas simultaneously bring on increased supplies, reduced demands for oil and gas, and increased use of alternative energy sources. In some sense the model results should reproduce observations in the real world of the energy markets in the context of the U.S. domestic economy observed in experience. This theme is consistent with our criteria of verifiability or correspondence.

Fourth, there is the question of the relationship of the model to "reality." It is tempting to state the fourth criterion (validity) in such phrases as "the model ought to represent reality." Such a criterion works reasonably well for model airplanes or planetarium models of the solar system, because we can observe both the reality (real airplanes or real plants) and the model and identify some correspondences. But the reality of an "energy system" and "the economy" cannot be observed in such a simple manner. Instead, there are data banks and theories, usually but not always put together by the disciplines, and are at best major abstractions. It is reasonable to expect that the model results should agree reasonably well with the accepted data and theories, and if it does not, then some satisfactory explanation should be forthcoming, e.g., that the data are incorrect or incomplete or that the theories require modification. One should expect greater difficulty in achieving acceptance of the model by the disciplines if the basic theories of relationships are violated or if one is attempting to develop new or modified theory in the context of a model designed to resolve policy issues. To develop a model representing a market economy in the absence of accepted economic theory is to challenge the basis for the economics profession and use of the model will, therefore, have a much different effect than a model based on well accepted economic theory. Coherence is at least partially dependent upon the understanding of the disciplines best represented by the current body of theory. The model must be understandable or coherent with reality in this sense. This theme is consistent with our criteria of validity or coherence.

In the Texas National Energy Modeling Project we have considered this set of four criteria from the modeler's, the model assessor's and final user's point of view. The implication is that the modeler should use such criteria and further, that a responsible modeler would do so. The criteria are also useful from the viewpoint of the model assessor. MREFS is judged by our group on its workability characteristics with primarily a user's point of view of important practical problems; to a lesser extent the model was evaluated according to its ability to raise important issues or to help resolve important issues from an analyst's, interpreter's or advisor's point of view. We have paid particular attention to the apparent intended uses of the model (as evidenced by its historical applications). Validity was judged in our project from the analyst's and final user's points of view. Extensive comparisons were made with the current body of economic theory since the model attempts to represent the market economy. We also paid attention to MREFS's physical relationships of geology and engineering which also must be coherent with accepted perceptions of reality in those fields. Verification, though not documented by the modelers in this case, was judged on the part of model assessors in our project on the basis of performance based on our own operation of the model as compared with documented expectations of the model builders and from logical deductions based on the theoretical underpinnings of the modeling system.

EVALUATION STUDIES

TNEMP commissioned eleven studies which fall into four groups: (1) computer

operation of MREFS (study 1), (2) supply modeling (studies 2,3,4,5,6), (3) processing and transportation modeling (studies 7,8,9), and (4) demand modeling with macroeconomic interface modeling (studies 10,11). This set of studies necessarily deals with physical relationships in engineering and geology, industry and consumer economic behavior, and the economics of conditional market equilibrium since this is the domain of MREFS.

Since MREFS is a large and complex modeling system developed over several years, with many man years of effort and with the expenditure of several million dollars, it was not considered practical to evaluate every aspect of the model and the data. Therefore, the _initial_ study design focused on the evaluation of (1) the crude oil, natural gas and coal satellite models with hands-on operation on the computer, (2) the transportation, electric power and refinery (and synthetics) sector representations in the integrating LP model, (3) the macro-microeconomic interface in all its aspects and the specific formulation of the demand models with hands-on operation of the DRI/MREFS interface on the computer, and (4) the integrating model with hands-on operation. The 1977 National Energy Plan version of MREFS was transferred to Texas for purposes of the evaluation. Figure 2 shows the portions of the modeling system evaluated in this study. The NEP version of MREFS was "brought up" on the Texas A&M University Amdahl computer and the principal investigator in charge of the system was to provide computer and operating support for other members of the team. Evaluation of other supply data and related exogenous calculations including uranium for nuclear power production, solar and other new technology contributions on the supply side, data and programs for estimating conservation impacts on the demand side, and data and calculations specifying non-crude oil import levels were to be ignored in the evaluation.

A mid-course correction in the study implementation set priorities for the hands-on operation on the computer to emphasize first the testing of the oil and gas supply model behavior with impacts on the integrating model solutions, second, the operation of the coal supply model with impacts on the integrating model solutions and third, the operation of the system to test macro/micro model interface results and the impacts of errors or variation in demand model parameter estimates. During the October, 1978 National Advisory Board meeting, including discussions with EIA personnel, it was determined that the NEP version of MREFS could not be verified by either EIA or TNEMP nor could TNEMP operate it intelligently, due to poor and non-existent documentation. Therefore, it was agreed that TNEMP would focus instead on the 1978 Administrator's Annual Report version of the model. This version was accordingly transferred and operated on the Texas A&M computer for TNEMP evaluation purposes.

MAJOR CONCLUSIONS CONCERNING THE MIDRANGE ENERGY FORECASTING SYSTEM

This section summarizes the major conclusions about the extent to which MREFS and its major submodels satisfy the criteria for evaluation. The conclusions are, to the extent possible, kept within the context of the state of the art for large-scale models in general, and energy policy models in particular. The statement is also kept in the context of the purposes for which the modeling system exists as indicated by how it has been used, to the extent we were able to ascertain such uses. This section deals specifically with conclusions concerning the reliability of major model outputs used in energy policy applications. To the extent possible, conclusions are drawn concerning the extent to which major model outputs are a function of exogenous relative to endogenous variables, and thus the explanatory power of the models.

Figure 2

Simplified Overview of
Midrange Energy Forecasting System Components*
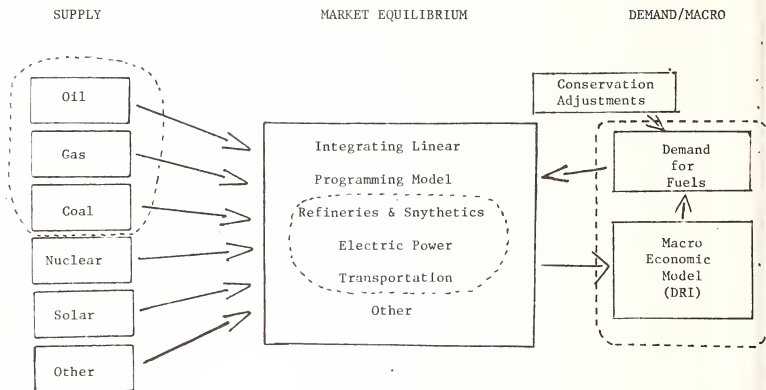
*Circled components indicate portions of MREFS evaluated in this study. The
oil, gas, and coal supply models were operated on the computer using selected
parameter variations for sensitivity analyses; the demand model was mechanically
operated on the computer but sensitivity analyses were done by ad hoc adjustments
to demand model parameters; existing results of the DRI model were used to test
the sensitivity of the macro economy impacts of energy market shocks; and the
integrating model was operated many times to incorporate pair-wise comparisons
of market equilibrium solutions from changes in the demand, oil, gas and coal
supply models.

MREFS Operations (Study 1)---. Study 1 deals primarily with the questions of clarity, workability and transferability of MREFS. This system is a complex combination of a number of formal models, data bases, parameter specifications and related software packages for operation and report writing. The need for the ability to transfer and operate large-scale models used in public policy analysis is crucial to the overall establishment of credibility and perhaps a major requirement for completing tests of validity, verifiability, workability and clarity.

The basic questions addressed in this study include (1) whether the MREFS can be successfully transferred and operated, (2) whether documentation is adequate for clarity, and (3) whether various simulations can be performed for purposes of validity, verification and workability tests. The study further provided for operation of successfully transferred models as a service function to the other principal investigators for individual study testing and operation.

The set of software systems installed and satisfactorily tested include:
1. Integrating Model (MEMM)
2. Regional Demand Forecasting System (RDFOR)
3. RDFOR Preprocessors
   a) Price Preprocessor
   b) Macro Preprocessor
   c) Parameter Preprocessor
   d) Initial Prices Preprocessor
   e) Conservation Shifts Preprocessor
4. Midrange Coal Supply Model
   a) does not include the entire National Coal Model (NCM)
   b) two University of Houston Fortran Modules have replaced the need for the Gamma Matrix Generator and Report Writer
5. Midterm Oil and Gas Supply Model System
6. Standard Integrating Table Preprocessors
   a) Oil Preprocessor
   b) Gas Preprocessor.

The various systems were installed and results checked against DOE/EIA results from the same systems operated on the DOE computer in Washington, D.C. Minor differences in the DOE check results and those produced at Texas A&M have been attributed to the following:
1. Version 1 of MAGEN used at TAMU versus Version 2 of MAGEN used at DOE
2. MPSX/360 used at TAMU versus MPS3 used at DOE
3. Different starting basis
4. Possible different alternate solutions
5. Replacement modules used in the Coal Supply Module.

The check results were exact except for the integrating system and the coal supply model. The base coal supply file produced by the modified coal supply model was input via the coal preprocessor and the final "cookie" (final report writer results) results were very close to the base "cookie" produced with DOE's AAR77 standard coal supply table. The differences in the integrating model results are in general quite acceptable. Some question has arisen in the area of non-associated gas results. DOE personnel suggested the possibility of alternate solutions when this was discussed with them. Some differences have been found in the output of Version 1 of MAGEN but this is not of major concern.

The major difficulty of implementation has been the incompatibility of the MVS operating system used at DOE versus the SVS system used at Texas A&M University. Many job streams had to be broken down into multiple jobs in order to prevent various queues from overflowing. Secondary difficulties were different physical intermediate I/O devices, lack of DASD allocation, required in or out specifications on the label parameter, and the inability to rewind DD * data files.

Many scenarios were run at TAMU in order to test the sensitivity of the overall model to major changes in selected parameters. The demand model was analyzed by initially varying the elasticities produced from the Regional Demand Forecasting Model. The consumption levels for fuels changed very little. The implication is that price paths which drive the model and are input to the model (predetermined) must be changed to significantly affect demand. This result was not clear at the outset.

Variations were performed on key parameters of the oil and gas supply models including (1) finding rates, (2) discount rates, (3) price paths, and (4) rig life. The variations were completed satisfactorily to allow analysis of the importance of these parameters on oil and gas production response behavior of the models.

Variations were also done on the "mine life" parameter of the coal supply model. These variations were also completed satisfactorily allowing analysis of coal production response behavior of the model.

The oil, gas, coal and demand model variations were interfaced pairwise with the integrating model in order to test the sensitivity of the overall market equilibrium solutions to the parameter variations.

One case (low finding rate for oil and gas) produced an unbounded solution problem in the integrating model. This occurred when attempting to insert the oil and gas supply curves corresponding to the 50% lower finding rates. Without the aid of DOE, this problem would have required an enormous amount of time to solve; one must thoroughly understand the workings of the model if such problems are to be solved. This was a new problem never encountered by DOE implying the need for more verification tests by DOE or others using the model. A "fix" is now being tested.

MREFS is definitely not of the "push button" variety and requires great attention to detail in the preparation of each scenario. However, the overall operation is quite straightforward and should be mastered by most reasonably proficient system analysts.

Supply Studies (Studies 2,3,4,5,6)---. Three factors are important to supply modeling and projection work, especially in the case of the oil and gas resources. First, the estimated size of the resource base that can be economically recovered (economic reserves) given current prices and technology must be known. Second, the physical relationship between drilling on the intensive and extensive margins and added reserves (finding rate) must be understood and estimated. Third, the behavior of the industry must be understood concerning investment, drilling and production decisions in response to expected prices of oil and gas and changing expected costs of drilling and production. The same basic factors are required for model-ing coal production; i.e., one must know the estimated economic reserves, physical recovery rates and economic behavior of the industry.

Study 2 deals primarily with the questions of validity and verification in the EIA oil and gas model. The evaluation provides the basis for showing the importance of estimates of economic reserves and finding rates on future production levels. The methodology used by EIA in projecting the finding rate for oil and gas, as the number of wells drilled increase the remaining resources decrease, was evaluated to determine the adequacy of the time path specifications used in the projection work.

The basic question addressed in this study is whether or not empirical estimates of economic reserves and finding rates may be of such large variation as to cause significant variations in output of the models regarding possible future production levels.

Two major conclusions summarize the study. First, the resource volume numbers, even given the uncertainty of undiscovered resource estimates, should not be a significant constraint in short- and intermediate-term projections (through 1985); it becomes critical, however, in long-range projections (1990 and longer). A reasonable variation in the resource numbers used by MREFS would not be expected, therefore, to _directly_ affect short- to intermediate-term projections. Second, historical data exhibit a distinct post-1970 exploration epoch, for which the finding rate has been relatively stable, and thus reliable for short-term (through 1985) projections. Analysis of regional shifts in the drilling mix show that these have very little effect on calculated finding rates. It is concluded that forward projections of reserve additions should utilize a finding rate in the range of 17 to 18 barrels of oil and gas equivalent per foot rather than the 21.5 boe per foot average implied in the National Energy Plan projections or the higher 24.7 boe per foot average utilized in EIA's Projection Series C through 1990.

The DOE calculation of finding rate, applying a continuous decline function to the past twenty years of historical data constrained by USGS reserve estimates, yields projected economic reserve additions 20 to 40% higher than recent history of finding rates would seem to warrant, and thereby underestimating the drilling effort necessary to support given production levels. The finding rate specification in MREFS is intuitively sound but is questionable empirically and lacks a valid basis for specification. Further research is needed to better identify the factors underlying observed finding rates for use in modeling. Study 5, which follows, summarizes MREFS behavior under varying finding rate specifications.

Study 3 deals primarily with questions of validity and verification in the EIA oil and gas supply models. By studying the investment decision framework of the EIA models and the current theory of investment behavior the study is able to make comparisons for the validity test. By studying empirically how the industry behaves the study provides direct and specific evidence of the oil and gas industry decision making process.

The basic questions addressed in this study are (1) whether the investment process depicted in the oil and gas supply models adequately reflect current investment theory, (2) whether industry specific empirical information supports or perhaps adds to understanding of the investment process, (3) whether empirical information from survey work can add reliability by incorporating empirical information into models of oil and gas supply, and (4) whether empirical evidence supports the idea that cash flow, in addition to marginal prices and other factors, influences investment decisions of oil and gas firms.

Four major conclusions characterize the study. Generally, the study

indicates a particularly serious problem is the apparent inconsistency of the investment model with capital budgeting theory and recent developments in risk measurement. First, the NPV (net present value) model was designed for the evaluation of a single investment project at the margin, not for determining regional drilling activity from average price and cost estimates as used in the oil and gas supply models. The regional approach averages out differences among individual firms and capital projects. This appears to introduce a bias in drilling activity at both high and low marginal prices. Alternative specifications of the investment process should be tested. Investment behavior models appearing in the recent econometric literature have generally relied upon both the demand for capital goods and the marginal capital supply function.

Second, the rate of discount assumed to motivate investment in the model appears to be seriously underestimated. A real discount rate of only 8% probably substantially underestimates the risk inherent in petroleum operations. A survey of petroleum companies suggests that a real discount rate of at least 15% after taxes may be more appropriate. The model was operated on the computer using a variety of discount rates to examine the consequences of changes in the cost of capital on the volume of drilling activity undertaken, and therefore on production. The model results for a change in the discount rate from 8% to 15% show an average (arc) elasticity (percent change in 1985 production from a 1% change in the discount rate) of -.578 for natural gas and -1.825 for crude oil at $1.75 and $12 prices respectively. The elasticities vary greatly over price ranges specified, ranging from -3.286 for gas at $1.00 per mcf to -.166 for gas at $3.00 per mcf and from -3.286 for oil at $10 per barrel to -.497 for oil at $16 per barrel. Further sensitivity results are summarized under Study 5 below.

Third, probably the most serious shortcoming of the model from the standpoint of financial theory is its failure to incorporate risk. To the extent that investment projects adopted by the industry add significantly to the total risk complexion of the firm, the model would tend to overestimate the volume of petroleum investment spending.

Fourth, survey results from a sample of independent drilling companies suggest it is inappropriate to assume that all petroleum firms use the NPV approach in making their investment decisions. A variety of approaches are used in practice and there is evidence that petroleum firms make use of NPV, pay-back, internal rate of return, and accounting-based methods in combination.

The survey results also suggest the critical importance of marginal prices on new output in determining petroleum investment. Net cash flow directly affects the industry's marginal cost of capital which influences its investment. Revenue generated by current investment directly affects industry sources of financing through an impact on the availability of internally generated funds. This circular relationship between cash flows, cost of capital, and investment suggests the necessity for a multi-equation framework within the model to more realistically reflect petroleum's capital-budgeting process.

The industry's increasing use of debt as shown in recent published data has generated the potential for greater volatility in returns to stockholders. This factor would suggest either that the discount rate applied to expected cash streams will have to be increased or the variability of estimated cash

flows (risk) included formally in the model if it is to more accurately mirror petroleum investment decisions.

Study 4 is primarily concerned with the question of validity in the EIA oil and gas supply model. By studying the behavior of the industry and examining secondary historical data on stripper well production the project determines whether special consideration should be given this class of producers in models which attempt to project oil and gas supply response to price changes in the current complex pricing structure maintained by the federal government.

The basic questions addressed in this study are (1) whether the peculiar behavior of stripper well owners warrants treating this class of producers uniquely in oil and gas supply models, (2) whether the growing population of stripper wells (currently 73% of U.S. total oil wells constituting 14% of U.S. total oil production) is likely to make this consideration more important over time, and (3) whether empirical data can be obtained for modeling this sector of oil and gas production.

The study results can be summarized in one major conclusion. The oil industry in MREFS is treated as a monolith. The stripper well segment is aggregated into the rest of the industry. The critical abandonment-produce decision that is continually faced by the marginal producer is assumed away when the exploration drilling decision is made in the model. In short, MREFS does not appear to model the economic characteristics of the stripper well industry and this part of the industry is of growing importance. Further study and data gathering will have to be done to incorporate this sector into oil supply models.

Study 5 deals with questions of validity, verification, workability and clarity in the EIA oil and gas supply models. By having hands-on operation of the models the studies are concerned with the actual behavior of the models, testing the sensitivity of major output variables to endogenous and exogenous variables.

By performing such tests and by examining the actual equation structure of the models the study determines the extent to which the specification conforms to economic theory. By having hands-on experience with the models the analysts were in a good position to assess the criteria of clarity. The study was also able to press toward understanding the requirements for verification.

The basic questions addressed in this study are (1) whether documentation and other descriptive materials are adequate for a user to operate and understand the models (clarity), (2) whether basic functional relationships can be synthesized for comparison with economic theory (validity), (3) whether the models answer the questions of production levels, drilling activity, and reserve additions needed by the user (workability), and (4) whether it is possible to provide tests of verification on these models.

Several highlights stand out from the evaluation. First, regarding validity, the model represents an attempt to be consistent with the theory of economic allocation of (drilling and production) resources, subject to declining finding rates and increasing costs of drilling. However, the fundamental joint product relationship between crude oil and natural gas supplies has been totally ignored. Ignoring this joint product relationship significantly undermines the validity of the oil/gas supply models because drilling capacity in reality may be used in drilling for oil or for gas alternatively, and further, some combination of

oil and gas is generally always found in every discovery. The result is to significantly bias the natural gas supply estimates downward and the crude oil supply estimates upward. This weakness raises a major question as to the validity of the model.

Second, use of a statistical finding rate estimate constrained by the U.S. Geological Survey's estimate of reserves significantly modifies the projections of finding rates from that of an unconstrained statistical estimate. There is no consistency between the economics of the model itself and the implied economics embedded in the USGS economic reserve estimates. This constraint means the economics of reserve additions are totally confused by the reserve estimating procedures of the Geological Survey. Institutional jurisdictions and responsibilities (reserve estimates by the USGS) have seemingly forced the imposition of a constraint, which may have nothing at all to do with the economics of additional oil and gas supplies. But this constraint has clearly modified the finding rate projections used in the oil and gas supply models, the supply projections made by the models, and the economic results obtained from the integrating model. The finding rate specification procedure in the model is conceptually nice but needs empirical content. That is, more empirical work is needed here to explain the finding rate behavior. No variations on this parameter are commonly shown with the EIA model results even though this factor is of obvious and great importance in production projections. For example, the behavior of the gas model when operated on the computer by analysts in Study 5 shows very little production response to price changes for prices near the current market price for the base case (parameter specifications by EIA for the 1977 AAR version of the model). Production responses in the gas model to marginal price changes show that for the base case production is very insensitive to price changes in the $1.75 to $3.00 range, decreasing drastically from that in the $1.00 to $1.75 range. For the lower finding rate case (50% decrease), however, production response to price changes is shown to be large even at high prices. Thus, the finding rate specification (or USGS reserve estimates) by which the finding rate is constrained) greatly affects the price responsiveness behavior of the model. The specification of finding rate also has a large direct impact on expected production levels at a given price set over the entire price range tested. For example, the elasticity of production with respect to finding rate for gas ranges from 1.131 for high finding rate/$3.00 gas to 3.000 for low finding rate/$1.00 gas in 1985. These findings have significant policy implications since (1) there is a considerable question as to the proper finding rate specification (see Study 2) with strong empirical indication that the EIA rate is too high by 20-40%, and (2) the model has been used in national energy debates to show that gas production response to price is very small at prices above $1.75. This model result needs further attention before the model is used for other policy analyses. The model suffers on the verification test.

Third, on the matter of clarity, the basic conceptual framework, once discovered, is reasonably straightforward. In implementation however, the model is massive, complex and convoluted. This situation, combined with incomplete (and occasionally erroneous) documentation, renders comprehension of the model unnecessarily difficult. Simply stated, the documentation would fail any reasonable test of clarity by the most understanding modeler.

Fourth, with respect to workability from the modeler's point of view, MREFS' oil and gas supply models are cumbersome and difficult to operate, their efficiency and flexibility suffering greatly from the model's patchwork origins

from the National Petroleum Council model. From a user's point of view the limited purpose model would be strained to answer more detailed questions about petroleum industry financial behavior or the impacts of complex regulatory measures now placed upon the industry.

Study 6 deals with questions of validity, verification, workability and clarity in the EIA coal supply model. By having hands-on operation of the model on the computer the study is concerned with the actual behavior of the models, testing the sensitivity of major output variables to endogenous and exogenous variables.

By performing such tests and by examining the actual equation structure of the models the study determines the extent to which the specification conforms to economic theory. By having hands-on experience with the models the analysts were in a good position to assess the criteria of clarity. The study was also able to press toward understanding the requirements for verification.

The basic questions addressed in this study are (1) whether documentation and other descriptive materials are adequate for a user to operate and understand the models (clarity), (2) whether basic functional relationships can be synthesized for comparison with economic theory (validity), (3) whether the models answer the questions of production levels and regional distribution of production from alternative demand scenarios, and (4) whether it is possible to provide tests of specification on the model.

Several highlights stand out from the evaluation of MREFS' coal supply model. First, the coal supply model is a valid representation of coal supplies in the intermediate term for which it was designed. In the long term the model has no provision for retiring "new" mines to account for the dynamics of coal depletion.

Second, concerning verification of the model, it adjusts faster than is possible in the industry because of lag times between building decisions and production start-up not accounted for in the model. The specification of one single lifetime for all mines in all regions needs verification with industry representatives. Mine life should be allowed to vary for different types of mines and different regions.

Operation of the model on the computer verified prior results by EIA. Variations in mine life assumptions indicated an apparently significant change in the supply functions for 1985 but the overall affects on market clearing prices from the integrating model solution was small.

Third, documentation of the coal supply model is reasonably adequate since the basic ideas have remained unchanged from the inception of MREFS (initially PIES). Improved clarity has been achieved at the University of Houston by eliminating one of the computer languages. From a user's or layman's point of view the model concept is easily understandable.

Fourth, workability of the coal supply model is adequate for its intended purpose of defining regional supply curves for MREFS' integrating model. The model will answer basic questions of relative regional coal production at various prices provided the transportation, electric utility and other coal-using modules in MREFS are correct.

<u>Process and Transportation Studies (Studies 7,8,9)</u>---. Three sectors are

important in transforming basic energy into usable energy forms and products for consumers and in transporting major energy raw material and products among regions of the country. This set of studies examines the MREFS representation of refineries (and synthetics), electric power generation, and transportation of basic forms of energy including coal, crude oil, refinery products and natural gas.

Study 7 deals with the questions of validity, verification, workability and clarity in MREFS' representation of the refinery sector and synthetic fuels production which is likely to be of importance as new technology develops. The representation of refinery sector operation including the change in the mix of products in response to changes in the relative price of products and cost of various crude types is examined.

The basic questions addressed in the study were (1) whether the model representation of the refinery sector corresponds to economic theory relevant to refinery operations (validity), (2) whether the physical relationships for converting crude oil into a range of products are maintained for a range of relative product prices (verification), (3) whether further development is needed to represent the synthetic fuels sector, and (4) whether the sector representations are capable of answering relevant questions within the context of the overall model (workability). The study amounted to an examination of the refinery sector representation in MREFS' integrating model including some computer runs to verify the behavior of the model.

Several conclusions summarize the study conclusions. The simple structure results in the total loss of technical structural delineations characteristic of the refining of petroleum products from various grades of crude oil, thus sacrificing the capability to analyze the consequences of structural change induced by energy policy. The model is not workable for policy questions related directly to refinery location decisions and other questions of the economics of the refinery sector; the model's rigidity and lack of environmental detail raises the possibility that the model might misrepresent the regional impact of policy changes on the refining industry.

Analysis of the model's behavior on the computer indicates that the model is "stable" only for a given range of product output/regional mix variations. One must know before hand whether the solution will be within bounds; the practical importance of this restriction is that these ranges are sufficient for short-term analyses but not for longer terms where refining equipment can be changed by the industry.

Study 8 deals with the questions of validity, verification, workability and clarity in MREFS' representation of the electric power sector. The representation of the electric power sector operation and behavior concerning the choice of fuels for generation, the capital structure, base load requirements, and other factors in response to changing rate structures and fuel costs are examined.

The basic questions addressed in this study were (1) whether the model representation of the electric power sector corresponds to current theory relevant to that of regulated utilities (validity), (2) whether the physical relationships concerning conversion efficiencies from various fuel mixes and changing base load requirements over time are maintained for a range of relative fuel mixes and growth rates for the sector output (verification), and (3)

226

whether the sector representation is capable of answering relevant questions within the context of the overall model (workability).

In the light of the evaluation criteria, the study makes three observations. First, the electric utility representation in the MREFS integrating model is the simplest possible. The only additional simplifications that could be made are to reduce the number of plant types or to reduce the number of load categories. Either would sacrifice appropriateness as an electric utility representation.

Second, the model will forecast behavior very poorly if unconstrained. However, the model has been used for the purpose of accounting; not for determination of behavior. Future expansion plans and plant utilization possibilities are entered as a very tightly constrained set of data. The model then is used simply to translate one set of prices (fuel and capital) into another (electricity).

Third, as an analytical tool in and of its own right, the model offers nothing that could not be obtained in more detail elsewhere. Thus the conclusion that the electricity sector of the MREFS model serves its intended, although somewhat limited, purpose adequately.

Study 9 deals with the questions of validity, verification, workability and clarity in MREFS' representation of the various energy transportation modes. The representation of rail, barge and pipeline transportation networks for moving basic energy commodities is examined. The transportation system representation deals with movements of coal, crude oil, natural gas and refinery products at the wholesale level.

The basic questions addressed by this study include (1) whether the behavior of regulated transportation modes corresponds to current theory (validity), (2) whether physical relationships of capacity, time constraints on new capacity and required product flows are maintained over a range of national and regional variations in the mix of energy production and consumption (verification), and (3) whether the sector representations are capable of answering relevant questions within the context of the overall model (workability).

Several major conclusions summarize the study results. The regional supply-demand structure of MREFS, using centroids within regions to represent the weighted centers of activity of different types, is an acceptable method to use in modeling energy distribution in the United States. However, several of the regions, such as the refinery region for the south-southwestern U.S. and most demand regions, probably are too large. In the case of coal, it appears that an attempt was made to correct for inaccuracies in the model basically caused by large regions by constraining the model solution.

MREFS' transportation models use constant transport costs for projection purposes. This assumption probably overstates the cost of transporting by pipeline and limits the model's usefulness for identifying changes in transportation modes over time. The forecasting method should have considered explicitly the trends in real transport cost by mode, thus raising a question of validity.

The transportation models used in MREFS have not been verified either indirectly, through verifying the overall MREFS model energy flows, or directly,

through determining how well the models predict transportation costs for different material/model combinations.

Although the concepts used in the transportation models in MREFS are easily understood, many of the data inputs are inadequately documented, which considerably limits the full understanding of these models by users outside the U.S. Department of Energy. This lack of complete documentation, together with the lack of verification, either of the overall MREFS model or the specific transportation models, leads to the conclusion that MREFS cannot be relied upon by outside users at this time to evaluate important energy policies heavily influenced by transportation questions. High priority should be given to structuring the model in such a way that it can be verified using recent data. Also the transportation models should be reformulated to: (1) explicitly consider trends in transport costs by mode, (2) consider alternative methods of modeling coal transport among regions, possibly using additional demand nodes, and (3) use more definitive models for estimating costs of new or upgraded pipelines and rail lines.

Demand and Macro/Micro Interface Studies (Studies 10,11)---. Three important modeling areas are crucial to the representation of the energy sector (and to energy policy effects) within the context of national and regional economies. First, the aggregate growth in economic activity has an important effect on the growth in energy consumption. Second, energy policy and energy sector behavior may have important macroeconomic effects or at least important effects on selected regional economies and economic sectors. Third, the direct consumption response to energy price changes and the substitution among energy forms and non-energy products and factors of production resulting from changing relative prices of energy are of great importance. Two demand related studies were designed to assess the importance of these questions.

Study 10 deals primarily with the questions of validity, verifiability, workability and clarity of demand models estimated for MREFS. MREFS specifies detailed demand functions by region and type of fuel. Demand relationships are specified according to economic theory, parameters are estimated with statistical procedures using historical data, and the functions (between price and quantity demanded) are shifted over time in the dynamics of the model to correspond to the influences of economic growth.

The basic questions addressed by this study include (1) whether the demand model specifications adequately correspond to current theory (validity), (2) whether empirical parameter estimates are reliable (verifiability), and (3) whether the dynamics of demand growth corresponds in significant ways to empirical data outside the data used in parameter estimates (verifiability).

MREFS' demand model is composed of three submodels, all of which are econometric in nature. The first, the residential, commercial and industrial submodel, is the most complex of the three, as these sectors accounted for about 73% of national energy use in 1975. From a validity criteria point of view the resulting model specification seems plausible but one cannot be certain since the original model specification is not documented or available, nor can one trace the rationale of the modelers from start to finish in order to ascertain the reasonableness of procedures. Thus, the validity of the model is questionable. The ordinary measures of statistical validity of the model parameter estimates essential to sound principles of statistical estimation--t coefficients and standard errors--are not available with the results nor were they apparently

used by the modelers, hence there was no way to assess the stability or relia-
bility of individual parameter estimates for the variables chosen for the models.

Projections for this model were accomplished by modifying the estimating
equations. Some variables such as heating degree days were dropped and the
distributed lag format was also eliminated before projections were made. Thus
the projection equation estimates had to be bench marked to a certain year to
guarantee comparability with estimating equation results.

The second part of the demand model, the transportation submodel, was
structured according to end use of the fuel. Hence there were, for example,
equations for auto highway gasoline use, non-auto highway gasoline and diesel
fuel use, rail diesel fuel use and commercial jet fuel use. In all, twenty-
two equations were employed to estimate demand for fuel for the transportation
sector, and single equation estimations, using ordinary least squares with time
series data were the main statistical tools used. The use of endogenous
variables as explanatory variables in the estimating equation leads to regres-
sion bias in the parameter estimates and the off-diagonal elements of the
error convariance matrix should be examined closely to determine whether bias
exists in equations so constructed. The t-tests, R-squareds and Durbin-Watson
coefficients were presented for these models and they appear to be satisfactory.
Price and income elasticity estimates also appear to be reasonable. However,
information on the regional demand for fuel by the transportation sector was
not made available to the investigators. Thus no evaluation of the workability
of the model at the regional level was possible.

The third submodel, the demand for minor fuels, included the following
relations: (1) the demand for natural gas, liquid gas and coal in the raw
materials sector, (2) the commercial sector demand for asphalt and liquid gas,
(3) the industrial demand for petroleum gases and metallurgical coal, and (4)
the residential and commercial demand for coal. Only relation (2) above in-
cluded price as an explanatory variable; all other relations used industrial
value added, time and lagged endogenous variables as independent variables. The
omission or price as an explanatory variable in these equations denies the basic
relation between price and quantity in demand theory. The use of time and
lagged endogenous variables seriously reduces the model's ability to correctly
estimate sudden changes in energy use that are quite different from past
trends, yet these types of changes could have been present in the middle
seventies. Statistical evaluation of the quality of parameter estimates was
not possible because of inadequate information provided to the investigators.

Sensitivity analyses were performed on the computer in order to determine
the practical importance of variations in price elasticity estimates in the
demand model. Several attempts were made to obtain reliable results from the
model system so as to determine if changes in the parameters have major impact
on the overall solutions from MREFS' integrating model. Instructions from
EIA personnel were followed after questionable results were obtained. The
intended sensitivity tests remain inconclusive.

Results obtained at this writing seem to imply the following; (1) target
year demand information (quantities demanded) for various end use fuels are
virtually fixed for the purposes of the integrating model solution, being
determined as they are by DRI macro variable forecasts and prior year's esti-
mates of "independent" variables through the lag structure of the demand
model. That is, so far as the demand side of MREFS is concerned, 1985 (or

1990) overall integrating model solutions are determined by the <u>price path</u> of previous years for final fuel demand by fuel type by sector; market clearing prices for the integrating model in 1985 are not significantly affected by a 1985 period price quantity relationship. As a result the overall MREFS results are heavily determined <u>a priori</u> by specification of the price paths.

Study 11 deals with the questions of validity, verifiability, workability and clarity of the MREFS representation of the interactions between the energy sector and the macroeconomy. The representation of the interaction is in terms of (1) market clearing energy prices from MREFS to DRI which satisfy equilibrium conditions in MREFS' forecast years, and (2) macro growth forecasts of personal income, employment, value added and other macro variables from DRI to MREFS.

The basic questions addressed in this study are (1) whether the representation of interaction between the macro and energy models of MREFS correspond to current theory of macro and energy sector economics (validity), (2) whether the interaction produces dynamic stability in terms of market prices, effects on the macroeconomy and consistent subsequent behavior of the federal government concerning basic macroeconomic policy (verifiability), (3) whether the conclusions of small macro effects of energy policy and large energy effects of macro policy are verifiable empirically, and (4) whether sector and regional economic impacts of energy sector changes as measured by MREFS are reliable on theoretical and empirical grounds. The study examines the extent to which MREFS is able to answer important questions of regional and sectoral impacts of energy policy (workability).

Major conclusions can be summarized as follows. First, regarding the DRI-MREFS interface, the structure of the linkage is seriously inadequate. The formal links from MREFS to DRI are the wholesale price index for fuels and two detailed prices and quantities, ignoring much regional and sectoral detail. This simple linkage raises serious questions of validity and greatly affects the model's workability since one is apt to be misled about the economic effects of energy market changes. Major adjustments to both models are done judgmentally by the modelers. There may be substantial regional, sectoral or industrial effects of changing energy prices, but these effects are not readily discernible by the models. There may be considerable short-run dislocations at the national level that are not captured by the long-run equilibrium solutions, again limiting the models' workability from a user's point of view.

Second, with respect to energy-regional economy interface, regions develop in lock-step with the national economy. The system is too simplified to provide much meaningful regional analysis to policymakers concerned with regional costs and benefits of alternative energy policies. Workability is weak for regional analyses.

Third, with respect to energy-international economy interface, the attempts to model functional relationships for oil exporting and oil importing countries suffer serious shortcomings. These shortcomings cause much of the analysis to be judgmental, with specific results being dependent on predetermined values of critical variables.

Overview

--The Texas National Energy Modeling Project (TNEMP) has achieved a
major mechanical transfer of a recent version of the Midrange
Energy Forecasting Model (formerly known as the Project Independence
Evaluation System) maintained and used for energy projections and
policy analyses by DOE/EIA (and formerly by the Federal Energy
Administration). Major progress has also been made in the operation
of the computer models by TNEMP evaluators to allow assessment of
the model's structure, behavior and usefulness for energy policy
analyses. This model transfer could not have occurred nor could the
limited evaluations requiring model operation on the computer have
been accomplished without major cooperation from DOE/EIA. TNEMP
represents a major step forward in improvement of model transfer and
evaluation required to advance the credibility and usefulness of
large-scale models for public policy analyses.

History

--MREFS was originally designed primarily for national aggregate projec-
tions concerning the outlook for energy consumption, production, sub-
stitution possibilities and import levels for crude oil; the modeling
system was designed to allow the Federal Energy Administration to
evaluate the prospect for energy independence resulting from a combina-
tion of government policies and prospects for new technology develop-
ment. As an initial effort in the face of the "energy crisis," the
modeling system design, construction and implementation was a major
task that deserves high marks from the standpoint of comprehensive-
ness and conceptual design and usefulness at the time.

--The continuing pressure for analyses and for related updates and
further development have left the modeling system seriously lacking
in areas of verification and documentation. This current state of
affairs leaves the modeling system seriously lacking and raises
important questions about the appropriate institutional arrangement
for development and use of such modeling systems. The models can
only be perceived as "black boxes" open to major abuses and suspicion
of major abuses. To achieve credibility a major effort is required
to document the models and enhance the ability of outsiders to under-
stand the workings, specifications, data and parameter estimates.

--Many of the model deficiencies identified in this report were known
by the original designers and identified in the appendix to the first
national study completed with the model, known as Project Independence
Blueprint. A research agenda to correct the deficiencies was apparently
thought through but never executed due to personnel changes and several
reorganizations. EIA now has a number of model development contracts
in progress and program plans for model development, documentation,
verification and access have been written.

Major Findings

--Not surprisingly, given the history and setting described above, MREFS
(PIES) did not measure up well under TNEMP evaluation criteria defined
as: (1) workability (aids in addressing important practical problems),

231

(2) clarity (is relatively unambiguous for analysts and users), (3) validity (the components of the model interact among themselves in a manner consistent with accepted understanding of how the system being studies really behaves), and (4) verifiability (there is some practical way to demonstrate that the model lives up to the designer's intent). Perhaps other large models, given the current state of the art, would not fare well under such criteria either but we should not be content with the status quo.

--MREFS' results are highly dependent upon the diffuse judgment of many modelers within the Department of Energy. This means that outsiders must necessarily have considerable faith in the integrity and judgment of the modelers themselves, as well as the "objective scientific" performance of a computerized modeling system.

--Given the current state of complexity and inadequate documentation of MREFS, transferability cannot be a mechanical operation relying on current written documentation. Transferability requires major interaction between the modelers and the model assessors or users of the transferred modeling system. The complexity of the modeling systems, the lack of clarity in the concepts and operations of the models, and the poor or non-existent documentation and almost total lack of verification information makes this kind of interaction absolutely essential to accomplish transfer; without transferability of the model and outside testing credibility badly suffers.

--Major reforms are needed to increase the reliability of information from MREFS. More specifically, third party institutions should specify standards for model documentation and verification procedures. Still another third party institution should be responsible for periodic assessments of the model's overall reliability in terms of criteria similar to that used in this study. The current DOE/EIA plan for archival of MREFS at Argonne, though helpful, is not adequate to achieve transfer; a "living" model must be maintained.

The Models

--The oil and gas supply models are seriously weak on several grounds and EIA would be wise to replace the model as is presently planned. The development of a new model, however, should make use of several findings of TNEMP. First, the representation of the investment behavior of the oil and gas industry is currently simplistic with no apparent empirical base. A survey of the recent economics and financial literature, as well as survey results from TNEMP suggest that a more complex and empirically based investment model is needed. Second, the models treat oil and gas as separate products when in fact the two hydrocarbon sources are joint products in the production process. Third, the discount rate used in the model is only about one-half that currently representative for the industry and considered necessary to encourage investment. Fourth, the finding rate calculation in the model is arbitrarily defined as a smooth declining function of time constrained by the U.S. Geological Survey's estimates of reserves; empirical evidence suggests a functional form reflecting various plateaus as a function of time. Further research is needed to explain this observed behavior.

Given the uncertainty of this parameter, variations should be pro-
vided in analyses since this factor is obviously of great importance
in production projections. The behavior of the gas model shows very
little production responsiveness for prices near current market levels,
a model result which is empirically questionable. Verification of
such model behavior is badly needed. Fifth, the model documentation
is so poor and/or non-existent that it is virtually impossible for an
outside user to operate the model with confidence. The usefulness of
the model for practical purposes of policy analysis concerning wellhead
price controls, tax incentives, wellhead taxes and other policy options
is therefore poor. The model lacks workability.

--The coal supply model is comparatively better then the oil and gas
models on matters of clarity and related documentation. To meet the
criterion of verifiability, more work on the model needs to be done
since the results indicate a more rapid industry production response
than observed empirically.

--The integrating model representations of the electric power, refinery
and transportation sectors are reasonably satisfactory given their
limited purpose as "place holders" in the market equilibrium structure
of MREFS; the overall solution of the MREFS is influenced primarily
by the supply and demand specifications. There are certain exceptions.
For example, the refinery sector representation produces reasonable
results (i.e., reasonable prices for the mix of refinery products)
for only a rather narrow variation in product mixes thus limiting the
usefulness of MREFS for some interesting analyses. The transportation
model representation is weak on verification grounds when used for
regional analyses.

--The demand model, which is central to MREFS projections of national
energy consumption and estimates of the economy's ability to substi-
tute fuels, does not contain certain test statistics used by econo-
metricians to judge the reliability of the empirical parameter esti-
mates. One can only guess at reliability by comparing results with
other studies which do have available test statistics.

--The DRI macroeconomic model used to drive MREFS is considered a
"state of the art" model and the idea of interfacing DRI with a de-
tailed energy market model is an innovative approach. However, the
linkage between the macro and micro systems is very simplistic and
ignores important interrelationships. Major methodological problems
arise because of attempts to integrate several modeling techniques.
Major sectoral and regional shifts in energy-related investment
resulting from changes in OPEC prices or U.S. government policy has
virtually no impact on the macro model results. The effects of U.S.
energy sector changes on world trade and its affect on the value of
the dollar are not directly modeled. Thus, the overall modeling
system can only provide partial information and could be misleading
about the most frequently asked questions by energy policymakers:
(1) What is the effect on U.S. economic growth of alternative energy
policy programs? (2) How will rising crude oil imports affect the
U.S. balance of trade, balance of payments and value of the dollar?

--A major weakness of the modeling system concerns its usefulness for

regional impact analyses. The methodology does not capture the regional shifts in economic activity driven by changing investment patterns resulting from changing relative prices of energy or government policy affecting the relative costs of doing business. Model results also indicate this weakness. For this reason the model is not usable for some of the most important policy questions facing policymakers--diverse and uneven regional impacts.

Model Assessment Criteria

--There is a need to further clarify a taxonomy of evaluation criteria. This exercise is a prerequisite to a common understanding of what measures should be used to test the reliability of information from large-scale models. This project has added to the understanding of what is involved in third party model assessments and transfer. Without DOE/EIA cooperation and many hours of personal communications at the staff level, transfer of MREFS, though still incomplete, could not have been accomplished. This fact reinforces our conclusion that transferability must be more than a mechanical operation.

Continued Texas Work

--Texas should further develop its own modeling capability for assessing the regional impacts of national energy policy proposals for input into the decision making process; no other institutional entity can be expected to adequately complete such analyses. Such a modeling effort should be done taking full advantage of the results of TNEMP experience.

DISCUSSION

Dr. Nissen (Chase Manhattan Bank): Before last December I was the proprietor of the PIE's Model and I built the demand model in PIES. I thought I would make two specific responses to what Milt Holloway said. I generally agree with the assessment that he has presented. I think the regional weaknesses are very much more important for Milt's purposes or PIES purposes now than they were for the model in its original guise. In fact, for the project Independence Blueprint version, which was the first version completed in 1974, the demand model operated at the national level. The totals were shared out to regions according to a set of fixed historical shares so that the demand model did not even depend on regional transient income and population, let alone have any feedbacks between the energy developments and the regional developments. So there has been an evolution towards regionality. It is generally unresponsive and incapable of dealing with regional energy economy interactions as it is currently configured. Also, generally it is incapable of handling, in any systematic way, national energy economy interactions. That is a very hard problem and I agree that it is an unsatisfactory state of affairs. There interactions have turned out to be very much more important, both at the regional, national, and world level-- much more than anyone expected in 1974. If you look back and look at the sort of impact assessments about what was going to happen to the world economy, I agree with that.

I would like to make a technical response which I think has some functional importance to this group as to the way the add factors were used in comparing the use of add factors in the DRI Macro Model which drives the macro forecast for PIES. Add factors as they are used in PIES, and I think--let me check this--what you especially mean is in the way the demand model is reconfigured to handle various specifications. Is that what we are talking about?

Mr. Holloway: Yes.

Dr. Nissen: Okay, add factoring is a precise term and it is a technical term in macro modeling. What you do is you have a set of equations that in general don't predict current history. So what you do is you change the intercepts so that they start at what we know is true now in the monthly exercise, if you are Otto Eckstein and you are running the DRI model. Then you either leave those in or you let them decay according to some things so that the representation of the model decays back to the data base that it was estimated on. That is an agnostic policy--a modeling process that simply recognizes that a good estimator of today's state of the world is today's state of the world, but there is no further analysis to that.

In the narrow sense, when we wanted to understand how the world looked with and without conservation programs, we did very much more than simply sort of wire today into the model. There was a collection of offline models that were part of an integrated community of policy assessment models used in the policy assessment activity which we carried out in '74 and '76 and then again in the summer of '77. The models I have in mind are the Residential Sector Model, where they are levers that represent policy actions and process standards and so forth and the Commercial Sector Model, a set of models operated by ICF and EIA that, respectively, estimated impacts of regulatory and pricing provisions. There was a documented analysis behind these shifts and there was a great deal of worrying about how the representation of these shifts within the consistent specifications of the demands models and the rest of the model.

So if you wanted to do what we always did functionally, that is recognize that these modeling efforts were part of the policy analysis system which was in play at that time, then they were integrated but not formally integrated in a code sense. And that is not add factoring. That is what you do when you have lots of big models which have lots of detail specificity-- is that you construct many model representations in some sense that are representable between various pieces of code and you worry a lot about the internal consistency of those analyses. So that kind of add factoring, which is not add factoring but is noncomputerized model integration is, I think, always an essential part of any rich, sophisticated policy evaluation carried on inside the government. The modeling assessment process has to understand that and recognize what it is and its importance.

REFERENCES

Gass, Saul I., "Evaluation of Complex Models," published in <u>Computers and Operations Research</u>, Vol. 4, No. 1, Pergamon Press, Oxford, England, 1977.

Gass, Saul I., "A Procedure for the Evaluation of Complex Models," presented at the "First International Conference of Mathematical Modeling," August 29–September 1, 1977, St. Louis, Missouri, published in the proceedings of the conference.

Halter, A. N., "Strengths and Limitations of Models from a User's Viewpoint: Modeling the Modeler's Model," presented at the "Modeling Symposium: Building Modeling Bridges Over Information Gaps," Continuing Education Center, University of Houston, April 29, 1977.

Holloway, Milton L., <u>Texas National Energy Modeling Project: Volume I Project Summary</u>, Texas Energy Advisory Council, Austin, Texas, draft report, March, 1979.

The M.I.T. Model Assessment Group, "Independent Assessment of Energy Policy Models: Two Case Studies," Energy Laboratory in cooperation with The Center for Computational Research in Economics and Management Science, Massachusetts Institute of Technology, draft report, May 15, 1978.

Selvidge, Judith, "Panel Discussion – Management Audit of Quantitative Models," Northeast AIDS Conference, Washington, D.C., June 1–2, 1978.

ACRONYMS

DOE	Department of Energy
DRI	Data Resources, Inc.
EIA	Energy Information Administration
MEMM	Midterm Energy Market Model
MREFS	Midrange Energy Forecasting System
NCM	National Coal Model
NPV	Net Present Value
RDFOR	Regional Demand Forecasting Model
USGS	United States Geological Survey

ASSESSING WAYS TO IMPROVE THE

UTILITY OF LARGE-SCALE MODELS [1]

Saul I. Gass
College of Business and Management
University of Maryland
College Park, Maryland

INTRODUCTION

Previous studies sponsored by the Department of Defense, The National Science Foundation, The Government Accounting Office, and the National Bureau of Standards have raised questions concerning the utility of computer based models developed for use by agencies of the Federal Government [4], [5], [6], [10]. These reports and others contain ideas and suggestions for improving the development, management and maintenance of a model during its life cycle. Based on our analysis, we identified eighteen broad areas by which current modeling improvement ideas can be grouped. These areas are as follows:

1. Data collection and availability for model improvement
2. Standardized procedures for model developers
3. Model user training
4. Model documentation plan and guidelines
5. Definition of large-scale models
6. RFP statement of work for model development
7. Model verification and validation plan
8. Relationship between model user and developer
9. Phased management approach to model development
10. Government in-house model development
11. Model post-review panel
12. Model ongoing review panel
13. Upgrading of the Government contract officer's technical representatives (COTR's)
14. Financial and milestone model management review techniques
15. Central model clearinghouse
16. Government model testing, verification and validation center
17. Government modeling research center
18. Modeling forums of users and developers

Up to the time of our study, there had been few attempts to determine (a) which specific model improvement proposals would work and be accepted by the modeling community, (b) what modeling research activities should be supported, and (c) if the Government was to support such research, what priorities should be established? This study was an attempt to answer these questions.

It is estimated that about 80% of the models developed for non-DOD agencies are constructed by organizations external to the Government; the figure for DOD models is 55% [4], [5]. Thus, it is clear that any attempt to impose ideas or guidelines or standards for improving the utility of Government modeling must take into account the concerns and interests of the modeling community at large. This community includes Government sponsors, users and model builders; private contractors; non-profit organizations; and university researchers and grantees.

STUDY METHODOLOGY

1. The basic problem was "How to Improve the Utility of Large-Scale Computer-Based Models?" We determined why this was a problem, who was involved and prepared a problem description paper.

2. We next defined the eighteen model improvement areas (as previously listed). For these areas we identified alternatives open to the Government in its desire to improve model utility. These alternative were generated from other studies and surveys, and from discussions with model users and developers.

3. For each area we developed a "model improvement proposition." We then discussed each proposition in terms of arguments for and arguments against. For example, one proposal was "for the Government to develop suitable standards for model development and to require the developer to conform to these standards whenever appropriate." Another proposition was "to require the model developer to prepare a verification and validation test plan and to report the results of the test plan in a technical report." While a third proposition was "as a means of improving the development and utilization of models, this proposal is to establish a central model clearinghouse."

4. For each proposition we developed related statements that highlighted and specified certain aspects of the proposition. For example, the proposition that dealt with standardization had the following three statements:

•A joint Government/industry committee should be established to investigate what aspects of the Government's modeling projects can be subjected to some form of standardization. The committee report would include a statement as to the costs and benefits to be obtained by standardization of particular modeling elements.

●Any modeling standards put forth by the Government should be voluntary and their application be left to the negotiation process between the model sponsor and developer.

●The development of modeling standards for computer routines, programming languages, data formats, etc. is <u>not</u> in the best interest of the Government's modeling activities.

The propositions for verification and validation, and central model clearinghouse had the following statements, respectively:

●A detailed verification and validation test plan should be required of most modeling projects. The project reports should describe the results and their implications to the future use of the model. Exceptions to a detailed plan should be based on a model's complexity and proposed use.

●The Government should establish a central model clearinghouse that would be responsible for the collection and dissemination of model documentation and related materials.

5. We then packaged the problem definition, the propositions, the pro and con arguments and related 32 statements into a review document. Fifty-seven modelers were selected to review the document. They were asked to express their views in three ways:

a. For the propositions that corresponded to the eighteen general areas that offered hope for improving model utility, select two high priority propositions and one low priority proposition.

b. For each of the 32 statements, express support or opposition to the statement on a five division scale: strongly support, support, undecided, oppose, strongly oppose.

c. Write comments in each area and in general.

In selecting the reviewers, we wanted to make sure that we had a cross section of modelers by affiliation and professional expertise. The affiliation categories were university, not-for-profit, profit and Government; the professional expertise categories were analytic, simulation and economics. The 57 modelers were selected based on their recognized expertise and their interest in advancing the modeling profession. Thirty-nine completed documents were returned. The final reviewer characteristics were as follows:

AFFILIATION:	University	8
	Not-For-Profit	8
	Profit	9
	Government	14
		39
EXPERTISE:	Analytic	19
	Simulation	12
	Economics	9
		39

STUDY DATA

The analysis of the responses was based mainly on the priority counts for each proposition and the support vs. opposition counts for each statement. The support or opposition counts were obtained, respectively, by a simple, unweighted sum of the strongly support plus support counts and the strongly oppose plus oppose counts. The actual uncombined counts and reviewer comments were used to interpret these combined counts.

For completeness, we next state each proposition and its priority count, followed by the associated statement(s) and combined support vs. opposition counts. We then offer an interpretation of this data. The proposition priority counts are given by the number couple (h to l), where h is the number of reviewers that denoted the proposition as being of a high priority, and l is the number of reviewers that denoted it as being of a low priority. The statement support and opposition counts are given by the number triple (s - u - o), where s is the total strongly support plus support count for the statement, u is the undecided count and o is the total strongly oppose plus oppose counts. Complete data by reviewer affiliation and professional expertise are given in the report cited in footnote 1. The information that follows is summarized at the end of this section in Table 1.

1: Data Collection and Availability for Model Development

The related proposition is that the scope-of-work should require a parallel effort in data collection either by the model developer, or other designated group, e.g., separate contractor, government agency. The scope-of-work should call for checkpoints that lead to specific decisions by the user and developer to continue the project based upon the status of the data effort. Prior to the RFP, projects should undergo a preliminary data availability and costing assessment. This assessment should be used by the sponsoring agency to continue or stop the effort. This proposition was third in terms of the high priority count (10 to 0). No university reviewer checked it for high priority.

Statement 1.1 Prior to the issuance of an RFP, most modeling projects should undergo a preliminary data availability and costing assessment where this assessment would be used by the sponsor to continue or stop the effort. Statement 1.1 was well supported by all categories of reviewers with university, not-for-profit and economic showing weaker support (31-2-6).

240

Statement 1.2 The RFP should require an explicit data collection effort to be conducted by the model developer or other designated group. Statement 1.2 has similar responses as 1.1 but its support was weaker, especially by the Government reviewers (27-3-9).

Statement 1.3 The availability of suitable data, as measured at certain milestones, should be a basis by which the sponsor and developer determine whether or not the modeling project objectives can be attained. Statement 1.3 has more opposition from the Government and analytic groups, but overall support is good (28-2-9).

Standardization

The related proposition is that the Government develop suitable standards for model development and require the developer to conform to these standards whenever appropriate. This proposition's priority count was (2 to 3).

Statement 2.1 A joint Government/industry committee should be established to investigate what aspects of the Government's modeling projects can be subjected to some form of standardization. The committee report would include a statement as to the costs and benefits to be obtained by standardization of particular modeling elements. Statement 2.1 has no real strong support. The reviewer groups tended to be supportive, except for economics (23-3-13).

Statement 2.2 Any modeling standards put forth by the Government should be voluntary and their application be left to the negotiation process between the model sponsor and developer. Statement 2.2 has very strong support from university, profit and economic categories (26-3-10).

Statement 2.3 The development of modeling standards for computer routines, programming languages, data formats, etc. is not in the best interest of the Government's modeling activities. Statement 2.3 (a negative statement) has more opposition from Government and simulation respondents. Profit and economic reviewers were more supportive (11-7-21).

User Training

The related proposition is that all model development contracts should explicitly address the training issue. If formal training is required by the contract, then training should include how to use and analyze the results of the model, along with data maintenance and program change procedures. If training is not required, the model specifications should state why it is not needed. This proposition's priority count was (3 to 0). These votes came from external Government, not-for-profit and profit model developers.

Statement 3.1   All modeling projects should address the need
to train others in the use and maintenance of a model, and where
appropriate, a formal training activity should be made part of
the developer's scope of work.   Statement 3.1 was well supported
by all categories (31-1-7).

4.   Documentation Plan and Guidelines

The related proposition is that computer model developers
specify a documentation plan at the beginning of the project and
have it approved by the Government Contract Officer's Technical
Representative (COTR), and that the Government develop documentation
guidelines for computer models, similar to the NBS FIPS PUB 38
guidelines for documentation of computer programs and automated
data systems [11].   This proposition received the second largest
high priority count (11 to 1).   It received strong priority
support from the categories of not-for-profit, analytic and
simulation.

Statement 4.1   As part of their contract, computer model
developers should specify a documentation plan at the beginning
of the project that details the documents to be produced, the
resources allocated and personnel responsibilities.   Statement 4.1
received the highest support count of all statements; this was
combined with the lowest opposition count.   It has very strong
support from Government, analytic and simulation categories (36-2-1).

Statement 4.2   The Government should establish a flexible
set of computer model documentation guidelines that can be
used by the model developer and the model sponsor to establish
a project's documentation plan.   Statement 4.2 received the second
highest support count.   It also has very strong support from
Government, analytic and simulation respondents (35-1-3).

5.   Definition of Large-Scale Models

The related proposition is that the Government develop a
basis for classifying models.   Any future modeling standards or
management procedures would then be applied based on the level of
a model's classification.   This proposition's priority count was
(2 to 3).

Statement 5.1   If modeling standards or management procedures
are developed by the Government, they should not be applied to all
modeling projects.   The basis for their application should be a
function of the project's characteristics.   Statement 5.1 has
rather strong support from all categories.   This support is weaker
for the profit and simulation reviewers (33-2-4).

Statement 5.2   The Government should develop a scheme by which modeling projects can be classified. The classification of a model would then serve as a guide to the model sponsor and developer as to the level and depth of certain modeling activities such as documentation, verification, sensitivity analyses, etc. Statement 5.2 has high total support count, but it is not very strong and is coupled with a fairly high opposition count (21-7-11).

6. RFP Statement of Work

The related proposition is that the RFP statement of work should indicate the technical and management aspects the developer must follow, including: specification of the analytical procedures, data to be used, reports required, and briefings that must be given. If a contract is to advance the state-of-the-art or a research area, then it should be so stated in the RFP and noted that model specifications are not being set at the start. The proposition priority count was (4 to 0).

Statement 6.1   The RFP should contain an explicit statement of the model's scope and objectives and indicate, when possible, the technical and management approaches to be employed by the contractor (model developer). Statement 6.1 has a high support count, with most of it coming from the support category. All reviewers support the statement, with university and economic reviewers showing less support (34-2-3).

Statement 6.2   If a model's scope and objectives are to advance the state-of-the-art or pursue a research direction, and specific technical aspects cannot be delineated in the RFP, then a procedure should be negotiated by which the model developer can ensure user or sponsor interaction during the developmental process. The interaction is a means for developing final model specifications. Statement 6.2 has a high support count from all categories, with a weakening of this support from university and economic reviewers. Very strong support from the analytic category (35-1-3).

Statement 6.3   An RFP statement of work for the development of a model should be reviewed by a special agency team (that can include external specialists) to determine if the model will be of value to the agency. Statement 6.3 has some support, little real strong support, and some opposition (22-6-11).

7. Verification and Validation

The related proposition is to require the model developer to prepare a verification and validation test plan and to report the results from this test plan in a technical report. The priority count was (6 to 0), with the priority support coming from the analytic and simulation professions, and the Government reviewers.

Statement 7.1 A detailed verification and validation test plan should be required of most modeling projects. The project reports should describe the results and their implications to the future use of the model. Exceptions to a detailed plan should be based on a model's complexity and proposed use. Statement 7.1 has strong support from all categories, with the university and economic support being diffused (30-3-6).

8. Relationship Between Model User and Developer

The related proposition is that the scope-of-work should stipulate who the ultimate user(s) will be, and that meetings between the developer and user(s) be held to enhance developer/user concurrence. This proposition had the largest high priority count (15 to 0), with a very high priority count from the analytic reviewers (10) and the smallest high priority count from the economic group (1).

Statement 8.1 Whenever possible, the ultimate user(s) of a model should be indicated in the RFP statement of work and the project plan require meetings between the model developer and user(s). The purpose of these meetings would be to aid the model developer in designing the model to meet user requirements. Statement 8.1 has very strong support from all categories, with the economic reviewers more diffuse in their support (32-3-4).

9. Phased Management Approach

The related proposition is that the Government further develop the idea of a phased approach similar in concept to the GAO proposal [10]. This proposition's priority count was (3 to 0).

Statement 9.1 The Government should investigate further the development of a phased approach to model management to determine the benefits and losses, what classes of models it could be applied to, and how it could be implemented. Statement 9.1 has a high support count, with most of it coming from the support category (rather than strongly support). It has support from all categories, with university and not-for-profit respondents showing more opposition (32-1-6).

Statement 9.2 The Government should develop some model management guidelines, phased approach or otherwise, that could be used by the model developer and model sponsor in developing and implementing their project plan. Such guidelines should be a product of a joint industry/Government group. Statement 9.2 has a high support count similar to 9.1 (31-4-4).

Statement 9.3 The development and use of a phased approach to modeling management is not in the best interest of the Government modeling activities. Statement 9.3 (negative statement) has a good opposition count, with a high undecided count. It has weaker opposition from university, not-for-profit and economic respondents (5-11-23).

10. In-House Model Development

The related proposition is to encourage more in-house research. More funds would be made available to support an agency's effort to do in-house model development, and probably less funds made available for contracting with an external developer. The priority count was (2 to 2). The tie vote is indicative of the statement votes in that they were the most indecisive.

Statement 10.1 The Government should attempt to increase the model development activities within Federal agencies, i.e., more models should be designed and developed by Government analysts. Statement 10.1 support and opposition counts were balanced off with a high undecided count (14-8-17).

Statement 10.2 The current balance between internal and external model development should be maintained. Statement 10.2 showed the same count pattern as 10.1, but it had the highest undecided count (10-17-12).

11. Post-Review Panel

The related proposition is to establish a post-review panel that would evaluate a model and provide guidance to potential users as to the model's strengths and weaknesses. The priority count was (0 to 2).

Statement 11.1 Each model sponsor should determine if the proposed model will undergo a post-review by a panel. If yes, the model sponsor, independent of the developer, should establish the review panel and the ground rules under which they will perform the evaluation. If no, the reasons why should be documented and become part of the project files. Statement 11.1 support and opposition counts were slightly in favor of support, with a low undecided count (21-1-17).

Statement 11.2 The decision whether a post-review panel will or will not be assigned to a model should be withheld from the model developer. Statement 11.2 has the highest opposition count of all statements and is opposed by all respondent categories (4-6-29).

12. Ongoing Review Panel

The related proposition is that a modeling project have an ongoing review panel. The priority count was (0 to 2).

Statement 12.1 Each model sponsor should determine if the proposed modeling project should have an ongoing review panel. If yes, then contractual requirements for meetings should be made between the model developer and the panel. Statement 12.1 has good support that is balanced by a high opposition count. The analytic reviewers showed no strong support in contrast to the simulation group that did (19-7-13).

13. Government COTR's

The related proposition is that the government should invest sufficient funds in training and development in order to employ and maintain an experienced and technically competent group of professional COTR employees. The priority count was (5 to 2), with profit reviewers having a priority count of (3 to 0).

Statement 13.1 The Government should initiate a program by which it would employ and train persons in the modeling profession who would then serve as COTR's for modeling projects. Statement 13.1 has about the same support and opposition counts, with a high undecided count. The Government group is basically neutral and the profit reviewers strongly supportive (16-10-13).

14. Financial and Milestone Review

The related proposition is to require all model developers to submit periodic status reports which allow the monitoring and comparison of accomplishments, personnel time, and selected expenditures against the original estimates made in the technical and business proposals. This proposition priority count was (1 to 2).

Statement 14.1 Modeling project contracts should require the model developer to submit periodic status reports that compare the project technical and financial plans to actual accomplishments. These reports would be used by the COTR to monitor better the progress of a project and to aid the developer in justifying any deviations. Statement 14.1 has a high support count. It is not enthusiastically endorsed, especially by the university and economic groups (25-5-9).

15. Central Model Clearinghouse

The related proposition is to establish a central model clearinghouse as a means of improving the development and utilization of models. The priority count was (2 to 5). This low priority count of 5 is the third largest.

Statement 15.1 The Government should establish a central model clearinghouse that would be responsible for the collection and dissemination of model documentation and related materials. Statement 15.1 has a higher opposition than support count. There does not appear to be any strong support or opposition by a group (14-6-19).

16. Model Testing, Verification, and Validation Center

The related proposition is that the Government establish a center at which certain classes of models would undergo verification and validation testing by an independent staff of analysts. The priority count was (3 to 15), with the 15 being the second largest low priority count. All groups rated it low.

Statement 16.1 The Government should establish a model testing center to which an agency may refer a model to undergo independent verification and validation. Statement 16.1 has a higher opposition count, with all groups except university registering strong opposition. The opposition count was the second largest (11-3-25).

17. Government Modeling Research Center

The related proposition is to create a Government Modeling Research Center. This proposition had the largest low priority count and was so selected by all categories (3 to 18). Most of the low priority count came from the analytic, Government and university groups.

Statement 17.1 The Government should investigate the need and value of a Government Modeling Research Center. Statement 17.1 has opposition from all groups, coupled with some support from all groups. Simulation showed the most support (14-3-22).

Statement 17.2 The setting up of Government Modeling Research Center is not in the best interests of the Government's modeling activities. Statement 17.2 (negative statement) has a high level of support from all groups, especially the university and analytic respondents (26-7-6).

18. Modeling Forums of Users and Developers

The related proposition is to organize forums like the Energy Forum in other key modeling areas. The priority count was (6 to 0).

Statement 18.1 The Government should establish modeling forums that deal with specific application areas and/or methodologies that are of concern to Government model sponsors and users. Statement 18.1 has strong support from all categories with the strongest from the university reviewers. There was no strong opposition (28-6-5).

Statement 18.2 Whenever possible, a modeling forum should be organized with the support of the appropriate professional organizations and industrial groups. Statement 18.2 has a very high support count from all categories and only one count in opposition (33-5-1).

TABLE 1

SUMMARY OF PROPOSITION PRIORITY AND SUPPORT/OPPOSITION COUNTS

PRIORITY	PROPOSITION
1. (10 - 0)	DATA COLLECTION AND AVAILABILITY FOR MODEL DEVELOPMENT 1.1 (31-2-6) 1.2 (27-3-9) 1.3 (28-2-9)
2. (2 - 3)	STANDARDIZED PROCEDURES FOR MODEL DEVELOPERS 2.1 (23-3-13) 2.2 (26-3-10) 2.3 (11-7-21)
3. (3 - 0)	MODEL USER TRAINING 3.1 (31-1-7)
4. (11 - 1)	MODEL DOCUMENTATION PLAN AND GUIDELINES 4.1 (36-2-1) 4.2 (35-1-3)
5. (2 - 3)	DEFINITION OF LARGE-SCALE MODELS 5.1 (33-2-4) 5.2(21-7-11)
6. (4 - 0)	RFP STATEMENT OF WORK FOR MODEL DEVELOPMENT 6.1 (34-2-3) 6.2 (35-1-3) 6.3(22-6-11)
7. (6 - 0)	MODEL VERIFICATION AND VALIDATION PLAN 7.1 (30-3-6)
8. (15 - 0)	RELATIONSHIP BETWEEN MODEL USER AND DEVELOPER 8.1 (32-3-4)
9. (3 - 0)	PHASED MANAGEMENT APPROACH TO MODEL DEVELOPMENT 9.1 (32-1-6) 9.2 (31-4-4) 9.3 (5-11-23)
10. (2 - 2)	GOVERNMENT IN-HOUSE MODEL DEVELOPMENT 10.1 (14-8-17) 10.2 (10-17-12)
11. (0 - 2)	MODEL POST-REVIEW PANEL 11.1 (21-1-17) 11.2(4-6-29)
12. (0 - 2)	MODEL ON-GOING REVIEW PANEL 12.1 (19-7-13)
13. (5 - 2)	UPGRADING OF THE GOVERNMENT CONTRACT OFFICER'S TECHNICAL REPRESENTATIVES (COTR'S) 13.1 (16-10-13)
14. (1 - 2)	FINANCIAL AND MILESTONE MODEL MANAGEMENT REVIEW TECHNI 14.1 (25-5-9)
15. (2 - 5)	CENTRAL MODEL CLEARINGHOUSE 15.1 (14-6-19)
16. (3 - 15). . . .	GOVERNMENT MODEL TESTING, VERIFICATION AND VALIDATION CENTER 16.1 (11-3-25)
17. (3 - 18). . . .	GOVERNMENT MODELING RESEARCH CENTER 17.1 (14-3-22) 17.2 (26-7-6)
18. (6 - 0)	MODELING FORUMS OF USERS AND DEVELOPERS 18.1 (28-6-5) 18.2 (33-5-1)

ANALYSIS AND CONCLUSIONS

A main purpose of this study was to develop a basis for discussing possible model guidelines, their feasibility and acceptability by the model developer and user groups. From the results of this study and our review of previous surveys and model research, certain directions are clear. An analysis of the available information should convince all interested parties that certain model improvement possibilities should be initiated, some dropped, and others put aside for the time being. Further research on improving model utility now must be focused on specific proposals. Future analysis should be directed towards the determination of the costs and effectiveness of these proposed activities. Then, the Government, in conjunction with the modeling community, should move to develop and implement the most beneficial activities.

To aid in the analysis, we have grouped the model utility proposals by six modeling activities. These are <u>model initiation</u> (propositions 4, 5, and 6), <u>model development</u> (propositions 1, 2, 8, and 12), <u>model implementation</u> (proposition 3), <u>model management</u> (propositions 9, 10, 13, and 14), <u>model assessment</u> (propositions 7, 11, and 16), and <u>model research</u> (propositions 15, 17, and 18). We next present our conclusions using these groupings.

A. <u>Model Initiation</u> (priority counts are in parentheses)
    Proposition 4: Model Documentation and Guidelines (11 to 1)
    Proposition 5: Definition of Large-Scale Models (2 to 3)
    Proposition 6: RFP Statement of Work for Model Development (4 to 0)

Conclusions: Proposition 4, dealing with model documentation plan and guidelines, should be selected for future development. If guidelines are developed, the results of proposition 5 indicate that they should be applied selectively, based on how the model is to be used. There is good support for proposition 6 and its statements on improving the RFP statement of work. Ways for doing this should be investigated.

B. <u>Model Development</u>
    Proposition 1: Data Collection and Availability for Model Development (10 to 0)
    Proposition 2: Standardized Procedures for Model Developers (2 to 3)
    Proposition 8: Relationship Between Model Users and Developers (15 to 0)
    Proposition 12: Model Ongoing Review Panel (0 to 2)

Conclusions: Propositions 1 calls for distinct data availability, collection and assessment tasks to be made part of a modeling project. This proposition should be selected for further development in terms of ways for improving the total data aspects of a project. There appears to be no real strong support for model standardization, proposition 2. Any standards should be directed towards computer aspects (languages, routines, structured programming techniques), but these should be voluntary. Procedures for strengthening the relationship between model developers and users are an overwhelming choice for development, proposition 8. In particular, formal meetings between the developer and user should be an RFP requirement. Proposition 12 calls for ongoing model review panels. It should not, in general, be pursued.

C. Model Implementation
    Proposition 3: Model User Training (3 to 0)

Conclusions: Proposition 3, dealing with the establishing
of training tasks, should be investigated at a future date. However,
major modeling projects that require the transfer of a modeling
system from developer to a Government user should incorporate
training and maintenance tasks in the contract.

D. Model Management

    Proposition 9: Phased Management Approach to Model Development
                    (3 to 0)
    Proposition 10: Government In-House Model Development (0 to 2)
    Proposition 13: Upgrading of the Government Contract Officer's
                    Technical Representative (COTR) (5 to 2)
    Proposition 14: Financial and Milestone Model Management
                    Review Techniques (1 to 2)

Conclusions: None of the model management proposals received
very strong support. However, proposition 9, on the phased approach
to model management, appears to have some good support from all
groups. We suggest that it be selected for future development,
depending on the availability of resources. An industry/Government
group established for the purpose of determining the form and
function of phased model management would be a low-cost way to
continue the investigation in this area.

E. Model Assessment
    Proposition 7: Model Verification and Validation Plan (6 to (
    Proposition 11: Model Post-Review Panel (0 to 2)
    Proposition 16: Government Model Testing, Verification and
                    Validation Center (3 to 15)

Conclusions: Model assessment is becoming an important
aspect of the modeling process. Proposition 7 requires a model
developer to prepare a verification and validation test plan and
to carry it out. This proposition should be further developed. Post-
review panels should not be implemented. Proposition 16 received
very little support and was the second lowest in priority. It
should be dropped from further consideration.

F. Model Research
    Proposition 15: Central Modeling Clearinghouse (2 to 5)
    Proposition 17: Government Modeling Research Center (3 to 18)
    Proposition 18: Modeling Forums of Users and Developers (6 to

Conclusions: The clearinghouse and modeling research center
proposal should be dropped from any future consideration. The center
received the largest low priority count of all proposals. The concept
of the modeling forum is endorsed strongly, especially forums
established by professional organizations and industrial groups.
The forum proposition should be selected for future development.

SUMMARY

Based on our analysis of the above material and reviewer comments, all categories of reviewers are against those model propositions that would tend to increase Government bureaucracy. The unfavorable reaction to propositions 15, 16 and 17 attests to this conclusion. However, there is recognition that the Government must begin efforts to improve the utility of the models it sponsors. Thus, strong support is given to model documentation plan and guidelines, and good support to model phased management and the RFP statement of work. At the same time, the responsibilities of the model developers are recognized in the support of the propositions dealing with verification and validation, and data collection and availability. The joint needs of the users and developers are recognized by the strong support of the user/developer interaction and the modeling forums.

We next summarize the above discussion as follows:

Propositions for Further Research

Strong Support

Proposition 1: Data Collection and Availability for Model Development (10 to 0)

Proposition 4: Model Documentation Plan and Guidelines (11 to 1)

Proposition 7: Model Verification and Validation Plan (6 to 0)

Proposition 8: Relationship Between Model User and Developer (15 to 0)

Proposition 18: Modeling Forums of Users and Developers (6 to 0)

Support

Proposition 6: RFP Statement of Work for Model Development (4 to 0)

Proposition 9: Phased Management Approach to Model Development (3 to 0)

Possible Support

Proposition 3: Model User Training (3 to 0)

Propositions Not To Be Supported for Further Research

Proposition 15: Central Model Clearinghouse (2 to 5)

Proposition 16: Government Model Testing, Verification and Validation (3 to 15)

Proposition 17: Government Modeling Research Center (3 to 18)

We feel that future activity on how to improve the utility of models should be concerned with the above strongly supported propositions. Ways for accomplishing these propositions, in a cost-effective manner, need to be explored and tested.

A final summation of the propositions and the reviewer information is, we feel, the following. Most of the supported propositions and their statements represent good modeling practices. It is not clear why these practices are not put to use on a regular basis. What the supported propositions call for is a better professional attitude toward modeling by all facets of the modeling community -- developers, users and sponsors. The reviewers have expressed how they believe the Government and industry modelers can improve the professional field of modeling and thus, increase the utility of models.

[1] "Advantages and Limitations of Computer Simulation in Decision-Makin
B-163074, U.S. GAO, Washington, D.C., May 3, 1973.

[2] "Computer Simulations, War Gaming, and Contract Studies," B-163074,
U.S. GAO, Washington, D.C., February 23, 1971.

[3] "Improvement Needed in Documenting Computer Systems," B-115369,
U.S. GAO, Washington, D.C., October 8, 1974.

[4] "Models, Simulations, and Games -- A Survey", M. Shubik and
G.D. Brewer, R. 1060-ARPA/RC, Rand Corporation, Santa Monica,
California, May 1972.

[5] "Federally Supported Mathematical Models: Survey and Analysis,"
G. Fromm, W.L. Hamilton and D.E. Hamilton, Stock No. 038-000-00221-0
U.S. GPO, Washington, D.C. 20402.

[6] "An Approach for Developing Feasible Guidelines for Large-Scale
Computer-Based Models," S.I. Gass, N.B.S. Report, P.O.T-73190,
March 1977.

[7] "Improvements Needed in Managing Automated Decisionmaking by Compute
Throughout the Federal Government," FGMSD-76-5, U.S. GAO, Washington
D.C., April 23, 1976.

[8] "Review of the 1974 Project Independence Evaluation System,"
OPA-76-20, U.S. GAO, Washington, D.C., April 21, 1976.

[9] "Auditing a Computer Model: A Case Study," Division of Financial
and General Management Studies, U.S. GAO, Washington, D.C., May 197?

[10] "Ways to Improve Management of Federally-Funded Computerized Models,
LCD-75-11, U.S. GAO, Washington, D.C., August 23, 1976.

[11] "Guidelines for Documentation of Computer Programs and Automated
Data Systems," Federal Information Processing Standards Publication
(FIPS) 38, National Bureau of Standards, Washington, D.C.,
February 15, 1976.

[12] "Guidelines for the Practice of Operations Research," The Operation
Research Society of America, Operation Research, Vol. 19, No. 5,
September 1971.

[13] "Evaluation of Complex Models," Saul I. Gass, Computers and Operatio
Research, Vol. 4, pp. 27-35, 1977.

[14] "Models in the Policy Process," M. Greenberger, M.A. Crenson and
B.L. Crissey, Russel Sage Foundation, N.Y., 1976.

[15] "Homilies for Humble Standards," D.T. Ross, CACM, Vol. 19, No. 11,
November 1976.

[16] "Requiem for Large-Scale Models," D.B. Lee, Jr., American Institute
of Planners Journal, May 1973.

] "Politicians, Bureaucrats and the Consultant," G.D. Brewer,
Basic Books, N.Y., 1973.

] "Systems Analysis in Public Policy," I.R. Hoos, U. of California
Press, Berkeley, California, 1972.

] "Large-Scale Models for Policy Evaluation," Peter House and J. McLeod,
John Wiley & Sons, N.Y., 1977.

] "A Procedure for the Evaluation of Complex Models," Saul I. Gass,
Proceedings, First International Conference on Mathematical
Modeling, August 29-September 1, 1977, University of Missouri-Rolla.

] "Evaluation of Policy Simulation Models," R.E. Pugh, Information
Resources Press, Washington, D.C., 1977.

VALIDITY AS A COMPOSITE MEASURE OF GOODNESS

Harvey J. Greenberg
Frederic Murphy

Office of Analysis Oversight and Access
Assistant Administrator for Applied Analysis
Energy Information Administration

HIERARCHY OF VALIDATION APPROACHES

Let us suppose validity is to be used as an evaluation criterion, expressing preferential selection among alternative models. We would like to say, "Model A is more valid than model B." This ranking could be confined to one application or to a spectrum of related applications. Using validity as a measure of goodness is consistent, for example, with the notion that validity, relative or absolute, pertains to a measure of confidence in the model, its output or its use.

The goal of this paper is to establish a minimum requirement for an objective evaluation of models. The evaluation is presumed to be based on the model's attributes. For example, an equilibrium model may have attributes: a supply function, a demand function, and a rule that defines a market equilibrium. Other attributes can be how these components interact to produce a forecast or the data bases used.

Today's economic models have many components, both structural and numerical. Such models are defined to be modular, although this term describes a form of model management rather than structure. While it is not necessary that an attribute be synonymous with a module or component, it is desirable to postulate a rule that relates a model's validity to the validities of its attributes. The reasons for this are twofold. First, the model validation process could be partitioned by component, easing the management task. Second, a "supermodel" could possibly be developed using the best design for each component of all models considered, eliminating the need for overall model comparisons.

The first question examined is whether it is sufficient to define the attributes for model evaluation as the validities of the submodels. We answer this question by showing that desirable properties of a concept of validity become internally inconsistent. Given the inability to analyze models piece-by-piece, even simple models become complex to evaluate, and the problem becomes one of multiattribute utility theory.

We then ask whether it is possible to retain some degree of simplicity in the validation process by using ordinal measures of utility. Given that we must consider both the submodels and their interactions, we shall demonstrate that Arrow's Possibility Theorem [1] can be applied to prove the logical inconsistency of properties that seem reasonable individually (see Fishburn's survey [2]). As a consequence,

if we postulate properties of an ordinal validity measure,
such as transitivity and completeness, and if we use a
"reasonable" rule about relating validity of models to their
attributes, then we encounter an inconsistency and must
forsake one of the properties, which we posited. Consequently,
there seems to be no reasonable simplification of the model
validation process that would allow us to judge models. We
are effectively left with the problem of comparing all aspects
of models, components, and interactions using cardinal measures
to express the degree of merit for every attribute. Since
the cardinal measures are dependent upon the scenarios of
interest and upon the evaluator, systematic model validation
for model selection remains a controversial task and may
be unattainable.

INITIAL PROPERTIES FOR VALIDITY

To address the issue of component evaluation, we need
only one of the four properties presented below. The four
are presented together to maintain the parallels with
Arrow [1]. Let X, Y, and Z denote three models, and let a
subscript denote an attribute--for example, X_i is attribute
i of model X.

The relation '<' is defined to be a validity relation,
where X<Y means "X is less valid than Y." We also define
X>Y to mean "X is more valid than Y," and X = Y means "X
and Y are equally valid." We use $X \leq Y$ ($X \geq Y$) to mean 'X<Y
(X>Y) or X = Y.'

The first two properties we shall assume are:

   Property 1 (completeness): For any two models, X and Y,
              one of the following relations is true:
                   X<Y, X = Y or X>Y.

   Property 2 (transitivity): If X, Y, and Z are three
              models such that X<(=,>)Y and $Y \leq (=, \geq) Z$,
              then X<(=,>)Z.

The validity relations (<, =, >), and associated properties,
also apply to attributes. When we do not wish to specify
the relation between X_i and Y_i, we write '$X_i R_i Y_i$.'

The Completeness Property (1) is axiomatic if validity is
to serve as a criterion for model selection. If, by con-
trast, the validities of two models are incomparable, then
validity cannot be the basis for choosing one of them.

The next property characterizes a relation between model
validity and attribute validity. It embodies a form of
coordinate-wise monotonicity. In particular, it says that
if one of the attributes is improved (in the sense that
$X'_i > X_i$) while others are left unchanged (i.e., $X'_j =
X_j$ for $j \neq i$), then the model's validity cannot worsen
(i.e., $X' \geq X$).

> Property 3 (monotonicity): If X and Y are two models
> such that $X_i \geq Y_i$ for all attributes (i),
> then $X \geq Y$.

The next property is Arrow's independence of irrelevant
alternatives. An example of what this property requires is
that it excludes the situation where the ordering of valid-
ities between two models, X and Y, depends upon whether a
third model, Z, is in the set of candidate models. Thus,
if X<Z<Y and model Z is deleted, it remains true that X<Y.

More generally, Property 4 requires the following to hold.
Assume X<Y<Z. Assume, also, the opinions on the validity
of the attributes of Z change relative to the attributes
of X and Y. (This change in attribute rankings may
change the ranking of Z relative to X and Y.) If we do
not change our opinions about the attributes of X, relative
to those of Y, then Property 4 says it must still hold that
X<Y.

> Property 4 (independence): Let (R_i) and (R'_i) be
> two equivalent attribute orderings for
> X and Y--e.g., $X_i<Y_i$ with R_i if and only
> if $X_i<Y_i$ with R'_i. If $X_1R_1Y_1,\ldots, X_nR_nY_n$
> implies X<Y, then $X_1R'_1Y_1,\ldots, X_nR'_nY_n$
> implies X<Y, irrespective of how (R_n)
> and (R'_n) ranks X_i and Y_i relative to
> Z_i, the i-th attribute of model Z.

There are two other properties that are reasonable to add.
Before presenting them, however, we shall illustrate how the
first four properties may be inconsistent if we presume a
component-by-component evaluation.

In this section we give an example where applying the monotonicity property exclusively to components may lead to a wrong conclusion. Although the first four properties were stated in ordinal terms, the definition of monotonicity would be unaltered with a cardinal scheme. The examples presented, therefore, represent the intrinsic inadequacy of component-wise measurements. First, to help affix ideas, consider an analogy in numerical error analysis.

Let B represent a computed inverse of a nonsingular matrix, A. Define the error matrix, $E = A^{-1} - B$. When we use B to solve a linear system, $Ax = b$, we compute Bb. The error is $Eb + e$, where e is the additional error from the computation of Bb. To keep notation simple, let e be negligible, so the error in computing $\bar{x} = Bb$, versus the actual solution, $x = A^{-1} b$, is $e(b) = ||\bar{x} - x|| = ||Eb||$. Obviously, it is desirable to make $||E||$ as small as possible.

There is a technique whereby one column of E can be reduced to zero, while leaving all other columns of E unchanged. That is, one component factor of the inverse can be improved while all others remain unchanged. This suggests the modified, computed inverse, B', induces less error than B. For some right-hand sides (b), however, previous cancellation of error is gone, and the solution error, $e'(b)$, is larger than the original error, $e(b)$. That is, we have $||E'|| < ||E||$, yet $e'(b) > e(b)$ for some b.

By analogy, we may have a model composed of three attributes: supply, demand, and equilibration. The equilibration component represents a model of market rules to arrive at a balanced forecast, thus embodying interaction between supply and demand. In its present state, we may know about imperfections in all three components. If we discover a way to improve one of the components, but leave the others unchanged, then for some scenarios (not predictable), the model results may have systematic biases not present before the "improvement."

For example, suppose we are interested in forecasting petroleum product prices. Assume we have an equilibrium solution from a supply model that underestimates supply for the given prices, a demand model that overestimates demand for the given prices, and a relatively inflexible refinery process model. These biases are not necessarily due to flawed design, but may be the consequence of concerns, such as model

size or the consequence of the technologies available to
the model builder. Assume we have the opportunity to use
a less precise demand model with overestimates of demand
cross-price elasticities and underestimates of total demand.
Using the criterion of least error in the estimated demand,
assume the first demand model is superior to the second.
A second model, consisting of the original supply and refinery
models, plus the second demand model, is likely to provide
the least biased forecast of prices; however, this is because
the overly large cross-elasticities compensate for the refinery
model's inflexibility in establishing refinery product prices.
Moreover, since demand is underestimated, we are less likely
to overestimate prices because of supply underestimates. On
the other hand, if we are more interested in quantities than
prices, then the first demand model is more appropriate.

MULTIATTRIBUTE UTILITY ASPECTS OF VALIDATION

We have shown that the logical consequence of a component-
by-component analysis has an undesirable result in that
there is no nonarbitrary way to arrive at a final evalu-
ation in a consistent fashion. The inconsistencies may be
avoided by adding an evaluation of model component inter-
actions. Thus, we must abandon complete modularity, but we
may consider other interpretations of attributes, such as
including key interactions of modules as attributes. Let
us explore further implications of maintaining the first
four properties by considering the two consequences that
correspond to Arrow's Possibility Theorem. Here we are
addressing the question of whether we can avoid cardinal
measures of attribute quality. Henceforth, we shall assume
that there at least three models and at least two
attributes.

The nondictatorial property, given below, says that no
attribute dominates the outcome when evaluating relative
validity. This, for example, eliminates lexicographic
orderings to determine relative validity of models from
their attributes.

The nonimposing property says that no two models have
an imposed ordering, irrespective of their attribute
orderings.

> Property 5 (nondictatorial): There does not exist
> an attribute (i) such that for all models,
> say X and Y, $X_i > Y_i$ implies $X > Y$ (no matter
> how the other attributes are ordered).

<u>Property 6 (nonimposing)</u>: There does not exist a pair
                  of models, say X and Y, such at X>Y for all
                  attribute orderings.

Arrow's theorem may now be restated in terms of the six
validity properties.

<u>Impossibility Theorem</u>

If properties 1-4 hold, then either property 5 or property
6 cannot hold.

The Impossibility Theorem raises the issue: Which property
should be discarded? The approach that was taken by econo-
mists, in the context of social choice, was to eliminate
transitivity. That may be appropriate in aggregating across
individuals, but it seems too important to drop in the val-
idation of models, because it says that the models linked
by the intransitivities really cannot be ranked. This would
violate the meaning of model selection on the basis of valid-
ity, which is the primary goal in defining validity as a
measure of goodness.

The independence property may be a candidate to eliminate.
For example, let us suppose that if there is a universal
truth, we shall never know it. Our references to "actual
data" to measure a model's forecasting accuracy, on which
we may build a validity ranking, really pertain to the
relative validity of <u>two</u> models, one of which is a measure-
ment model, which we <u>believe</u> to be the most accurate. In
the notation of the independence property, let Z represent
the measurement model which produces what we call actual
data. In this case an improvement in Z may well affect
the relative ranking of X and Y. If, however, the change
in Z causes us to re-evaluate X and Y, then we have some
underlying cardinal measure based on forecast deviation.
Indeed, independence is the axiom that is discarded when
a cardinal validity function is assumed.

AN EXAMPLE OF THE DIFFICULTIES OF MODEL COMPARISONS
USING A CARDINAL SCHEME

To illustrate the difficulties of establishing the relative
validity of two models, let us compare the embedded utilities
submodel in the Mid-Range Energy Forecasting System (MEFS),
formerly called PIES, with the Baughman-Joskow (B-J) model.

A reasonable, but superficial, view is that the Baughman-Joskow model is better able to forecast fuel use and electricity prices. More generally, in a short list of important attributes, the following comparisons are made:

	Preference
Financial Issues	B-J
Capital Acquisition	B-J
Dispatching	MEFS
Data	MEFS.

The first two preferences listed are based on the feature that the Baughman-Joskow model has more detail on financial aspects of utilities, including a focus on capital availability. Further, in the construction of their model, Baughman and Joskow emphasized the problem of representing demand and price expectations dynamically, along with the influence of expectations on equipment selection. The third preference is based on the different areas of emphasis of the two models. Since the MEFS utility model distinguishes new and existing equipment, it has greater disaggregation of operating characteristics, resulting in a better representation of dispatching decisions.

Finally, the last preference reflects the relative quality of data on capacities, heat rates, and other physical properties. (The data is probably superior in MEFS because of the large amount of resources that are continually devoted to the task.)

Since the capital acquisition methodology determines the equipment available for dispatching, the Baughman-Joskow model should be considered the better model. Nevertheless, because of utility lead times, new investment beyond the currently announced expansions is less than 12 percent of the 1985 rate base in one of the projections from the 1977 EIA Administrator's Annual Report. In this situation, given no unusual interruption in the demand profile, a better representation of existing equipment and dispatching is the dominant consideration to forecast utility behavior. When, however, the forecast is beyond 1995, the Baughman-Joskow model is asserted to be better because more of the existing equipment would be retired, and demand levels would be higher, making new equipment selection a more important factor in the forecast.

Although the comparisons presented between the two models are simplistic, they illustrate the level of difficulty in

cardinal evaluations that go beyond component-by-component
evaluation. Further, measurement of validity requires
knowledge, not only of model structure, but of the scenarios,
in order to provide a nonsuperficial comparison. In partic-
ular, we may believe that a scenario designed to answer
questions about mid-term impacts is best answered by MEFS,
because the important factors are scheduled capacity expan-
sion and dispatching, while long-term impacts are better
addressed by B-J because what is available to dispatch is
well beyond a data issue; the important factors affecting
capacity expansion involve financial considerations and
expectations that project into the distant future. A car-
dinal measure, therefore, must depend upon the horizon of
the forecast. For a horizon of 5 to 10 years, MEFS pro-
vides a better forecasting capability; as the horizon
moves into the more distant future, B-J is the more valid
model to use.

In addition to the temporal considerations, there may be
attributes overlooked due to unfamiliarity with the model.
Once we admit that validation is scenario-dependent, then
it must also be dependent upon the people who interpret the
model results for its use in analysis. We call these people
the "modelers," and they may not be the "evaluators" in a
validation exercise. We maintain, however, that evaluators
may be unable to rely on the validation process to judge
which is the better model to use for a specific study. In
other words, it is at least difficult to separate the modeler
from the validation process and still do a thorough job,
so total objectivity is not realizable without sacrificing
quality.

CONCLUSION

The problem of formally defining validity has a fundamental,
irreducible complexity. If a component-by-component com-
parison would be sufficient, in many cases there would be
none of the difficulties of aggregation. This is because
the best model could be created by combining the best compo-
nents of the candidate models. Since this was shown to violate
a "reasonable" property for the validation function--namely,
monotonicity--we are left with the position that all we can
say is, "This model represents this component better than
that model." The global effect of such superiority remains
uncertain, and the process of validation cannot be modular.

We also lose the option of developing an ordinal approach
to aggregating attributes and must ask for the "intensity"

of the rankings--that is, a cardinal measure. Since modelers are likely to disagree on the importance and the validity of each attribute, there is likely to be no consensus about the relative merits of a collection of models. Thus, peer assessment is in direct conflict with rigor and consistency.

We are left with an incomplete definition that includes only the Completeness and Transitivity properties. Further, we have shown that an ordinal scheme must necessarily lead to inconsistency, so a cardinal scheme must be developed. Peer assessment, with an explicit voting criteria, has an intrinsic inconsistency similar to the ordinal scheme. All that appears to remain is the hope that there could evolve a scheme to categorize models and scenarios such that an element of the induced "category matrix" may have a suitable cardinal scheme. For example, if we confine models to be econometric or statistical, and if we confine scenarios to be short-term, then forecasting accuracy, measured against "actual data," may be suitable.

DISCUSSION

Mr. McKay (Los Alamos): In your conceptualization of model validation, let's suppose you wanted to fit in the idea that if in the process of trying to validate or invalidate a model, you discover you don't like what you see. Do you feel that it is your duty or your responsibility to try and point out to the model developer the source of the trouble within the model? In other words, not merely to report this didn't work right, I didn't get a good enough answer, but I think it didn't because.

Dr. Murphy: I would go beyond that. I would say not only it didn't work because I would also say this is how I would fix it because I view the assessment process as one of not only model transparency, but one of model improvement as an insider.

Mr. McKay: I believe if you are trying to look at models, data and structure, and to decide whether or not the data is good enough or the structure is good enough, that it almost goes hand in hand with being able to identify or trace back incorrect answers from a model to the data or the structure. I have found that is a very difficult thing to try to do conceptually and formally, somewhat as you have done up there, and I was wondering if you had any ideas on that.

Dr. Murphy: Are you talking about the first part or the second part of the talk?

Mr. McKay: First part.

Dr. Murphy: The first part of the talk was really only trying to come up with a sensible organization for the validation process. Could we compartmentalize it on the various model components? The answer is no. Can we try and eliminate people's value judgments to degree as well as difference? The answer was no. So far as pragmatic model validation, I think that the general conclusion of the first part is that it is apparent to me that there is no simplification of the process. That doesn't mean that I'm going to convert that into a point of action, except to realize that you have to worry about all the complexity.

Dr. Wood: A few comments and a question Fred. It seems to me it is a bit strong that the MIT approach does not include invalidation correspondence of the model with theory. For example, I think a lot of our report is given over to the structural validity of that model. I think that there is also a strong emphasis, perhaps the greatest emphasis, on the validity of the model with respect to particular policy applications. So this notion that we tend to always be saying what is wrong with the model, it is important to put that in context. It is always in the context of a particular application. I think also that we pay attention to the other elements of validation. For example, in the discussion of the demand model, that is almost entirely a structural validation exercise in which we compare that model to economic theory and make inferences about what we would expect the model to look like, check it, and then verify that it, in fact, is implemented that way.

I found your discussion of how you are organizing it extremely provocative and I think that it is going to be interesting to see how you are going to thread your way between assessment and countermodeling. That is you are talking about constructing a counter model now which involves a different formulation of utility investment behavior of profit maximizing under a regulated rate of return, that is going to involve the new model. At the end of that process it will be interesting to see what you can say about validity and verification of the original model and how that relates to your counter modeling activities. That is what I am concerned about and is what Dave Kresge was talking about. Won't your objectives become a little mixed--on the one hand you are a scientist seeking knowledge on how utilities behave and on the other hand you are an analyst of this particular model attempting to validate and verify it and it seems to me that is a tricky road to walk down.

Mr. Murphy: Is that the question?

Dr. Wood: There is a question in that. The third thing I was going to say was I wondered how you reacted to--that is there are other people, Saul is one of them, that feel that you can structure scoring systems for comparing models and I was wondering what you thought about those propositions?

Mr. Murphy: The first thing I guess I should clarify my statement in that the MIT approach was to start in the model and then work out. The approach we are taking is let's not even look at the model to begin with

and let's ask what we ought to have--what are we looking for. Essentially, I am pursuing in part Hoff's approach. It says the assessment process ought to be parallel to the model building process because you want to know what you need first. So my approach is to start and ask what is the menu of things that you think you ought to have and then start asking does the model capture that? The MIT assessment started with--and again it is because of the circumstances--I know the PIES Utility model--you didn't know the Baughman-Joskow model--see, you went through and you did an assessment of what was there which then led you to look out, should this have been done differently? So that is the emphasis I meant to make.

Your second question was--I don't know that there can be a distinction in what should be the consequence of the assessment process. Should it just be information for third party judgments or should--a tremendous amount of funds are expended because model assessment costs about as much as model development--it be channeled to offering the positive suggestions to model improvements. In other words, the glass is half full rather than the glass is half empty philosophy and so I don't see any conflict there. I see that there is a possibility for the assessor to become stale if he lives within the world of his model too long, but I think that the consequence of the assessment process has to be improvement, or a statement that improvement is unnecessary. The only way to know if the improvement is unnecessary is to try the improvement and to measure the difference and see if the difference is worth the effort. The question, can you put a rating scheme on models? Well, if you go other than zero and one, or minus one plus one, and you go from zero to twenty including all the numbers in between or the integers in between, you have gone to a cardinal scheme. So what we said in the first part you can still be consistent. Essentially what happens when you go to a cardinal scheme is the property you give us is independent. George Lady has a very nice corollary to that axiom in showing really the extent to which that is--where the ordinality is the key requirement. As soon as you add more than two numbers you are cardinal.

REFERENCES AND BIBLIOGRAPHY*

[1]   Arrow, K.J., _Social Choice and Individual Values_,
        New York, Wiley, 1963.

[2]   Fishburn, P.C., "A Survey of Multiattribute/Multi-
        criterion Evaluation Theories," in Stanley Zionts
        (ed.), _Multiple Criteria Problem Solving_, Springer-
        Verlag, New York, New York, 1978.

[3]   Fishburn, P.C., "Lexicographic Orders Utilities and
        Decision Rules: A Survey," _Management Science_,
        Volume 20, No. 11, 1974 (pp. 1442-1471).

─────────────

*   See also the voluminous bibliography provided by
    Fishburn [2,3].

THE IMPACT OF ASSESSMENT ON THE MODELING PROCESS

David Nissen*

When Saul Gass invited me to present a paper to this conference, I welcomed the opportunity for two reasons. First, it offered a chance to organize my personal perspective on a very exciting and fruitful period of my own professional life. (In 1974-77, I participated in, and later directed, the Project Independence Evaluation System (PIES) modeling and policy analysis activity at the Federal Energy Administration.) Second, I could present for public scrutiny some hard-bought and, I hope, useful lessons drawn from that experience.

BACKGROUND

Energy modeling for policy analysis is a burgeoning industry by any standard. The PIES effort served as a constituent of this success, and as an example of the problems which success creates. PIES also served as a seed-irritant in the energy policy advocacy process, which set in motion forces leading to a focused and institutionalized concern with energy model validation and assessment. Our presence at this workshop is a consequence of that concern. To understand this concern and how to meet it, it is valuable to examine the context in which it evolved.

PIES was initially developed to coordinate the quantitative assessment of the Administration's response to the embargo and oil price run-up of 1973-74 and the changed energy perspective which these events induced.

At first, the modelers had to convince the immediate clients, their management, of the accuracy and relevance of the model, and of its responsiveness within the policy decision horizon. This occurred during, not after, development of the model, which meant the first level of users was unusually familiar with the innards of the model. (The point is that the decision to develop and use the model was itself a policy issue--it was expensive and risky and required a lot of interagency organization. The fine structure of a model in place could never have commanded or sustained this level of attention by the management on its own merit.)

*The author is Vice-President, Energy Economics, Chase Manhatten Bank, N.A. He is grateful for comments and criticisms to Edward Cazalet, Harvey Greenberg, William Hogan, David Knapp, George Lady, Fred Murphy, Lee Nissen, Warner North, James Sweeney, and James Wallace. The views expressed here are the author's and do not represent the position of any institution.

Because of the client/management's familiarity, modelers were asked to be, and were willing to be, much more adventurous in modeling scope—the breadth of phenomena and policy issues that were integrated into the model—than is the case in the more usual analyst/policy-maker relationship.

In other words, from the viewpoint of both the client/management and the analyst/modeler there was a high immediate payoff to model enhancement for new or more accurate and sophisticated policy evaluation while at the same time there was minimal immediate need for the more formal exegesis (including but not limited to documentation) that a more distant relationship between modeler and client requires for success.

It is not surprising that the allocation of resources within the modeling process reflected this emphasis on development, to the detriment of investment in formal external communication of the model's nature. This emphasis is apparent throughout the four major epochs of PIES' formal existence (the name and function of the model have changed under the present DOE management and organization). These are:

o   1974—construction of data and logic of the first version (the competitive equilibrium version) of PIES to produce the quantitative analysis for the Project Independence Report,

o   1975—extension of structure, including oil price-control modeling, and consolidation and extension of data to make the model reliable and robust for state of the world and policy scenario variations published in the 1976 National Energy Outlook,

o   1976—refined capability for policy analysis including gas regulation modeling (82 scenarios implementing a 50-page policy analysis specification) published but not disseminated in the 1977 National Energy Outlook (Draft),

o   1977—analysis of the National Energy Plan—adaptation of the model's structure to coordinate analysis of the conservation, fuel pricing and fuel management policy options being considered and advocated by the present administration, the results being published in the April 1977 white book, The National Energy Plan (Energy Policy and Planning, Executive Office of the President, April 29, 1979), and subsequent White House fact sheets and backup documentation.

By September of 1977, when the Carter Administration's National Energy Plan had reached the Senate (there to languish for a year) the education which the model could provide, and which the Administration was willing to absorb, was largely complete. The national debate on energy policy had been joined on larger questions of institutional means and of interclass and intergenerational equity which PIES couldn't begin to organize or resolve.

Throughout this four year period, the changing needs and goals of the model's internal clientele in FEA and the White House received paramount attention, often in the face of the modelers' plaintive objections, which reflected criteria both of professional workmanship and personal career goals.

In his invaluable book, The Mythical Man-Month (Addison-Wesley, 1975), Fredrick P. Brooks, Jr., "the father of IBM System/360," distinguishes four stages of a programming effort:

$$\text{Program} \xrightarrow{\text{x 3}} \text{Programming System}$$

$$\text{x 3} \qquad\qquad\qquad \text{x 3}$$

$$\underset{\text{Product}}{\text{Program}} \xrightarrow{\text{x 3}} \underset{\text{Product}}{\text{Programming Systems}}$$

Briefly, a <u>program</u> solves a problem for its authors on the system on which it was developed. A <u>programming system</u> solves a class of problems for its authors on the system on which it was developed. A <u>product</u> is a documented, debugged and transportable program or programming system which can be used to solve a well defined problem or class of problems by anybody on any suitable system. Brooks estimates that each development takes three times the effort of the previous one.

Computerized models are different from computer operating systems but the phylogeny is similar. Within this taxonomy, due to the sophistication and receptiveness of its internal clientele, the PIES development always moved towards developing a modeling system to enlarge its scope, at the cost of developing a model product.

The managers of PIES defend this choice as the strategic and efficient deployment of modeling in support of energy policy analysis, and I think they were right. Nevertheless, the present counter-revolution to crash-mode model development and use for policy analysis must be counted in the costs of this choice.

As a consequence then the modeling activity was a palpable force and resource in policy development and policy advocacy while at the same time its external relationships were not systematically developed. Perhaps inevitably a backlog of suspicion, resentment and fear of the model developed in the Congress, in other government agencies, and in private organizations with a stake in the energy policy process. Lack of public understanding of the model's structure, data conventions and data crippled counter-modeling efforts in the policy advocacy process. Thus, as long as the proprietors of PIES dictated the quantitative framework of the debate, PIES as a modeling facility could be used to bully the opposition with numbers.

In effect, PIES could never be assessed, let alone used, by those who didn't manage it. As a consequence, even if its answers were objective and

neutral (as we modelers' claim), its use in the policy advocacy process was not.

The response of the Congress and the new Administration to this situation, and parallel situations in data activities, has been to insulate the EIA management of the modeling and data activities from direct Administration control (in the process, relocating the responsibility for crash-mode policy analysis in the DOE policy function), and to establish formal organizations for assessing and monitoring these activities. These take the form of the Professional Audit Review Team (PART) externally, and the Office of Data Validation and the Office of Analysis Oversight internally.

To do their job, that is, to organize review, validation, and oversight operations, and to communicate their results to the Congress and the external audience, these organizations will, of logistical necessity, establish formal assessment procedures and validation standards.

The procedures and standards will profoundly affect the way modeling is done. These impacts of assessment on modeling (I include validation in assessment) are the subject of this paper.

OVERVIEW

My thesis is that the development of modeling assessment procedures and validation criteria is a natural and inevitable consequence of modeling success in policy advocacy and evaluation. Further, this development can be economic in that it will facilitate more widespread and effective use of the modeling activity in the production of policy evaluation. The role of assessment and validation can be understood as part of the "economic development of the modeling industry" literally construed. Assessment and validation activities provide the context in which will develop the extension of markets, the specialization of functions and the elaboration of intermediate products which are the concomitants of economic development. Finally, identification of assessment as the market infrastructure in which model development occurs allows identification of the natural and desirable impacts of assessment on the modeling process. These are enhanced emphasis on modularization, standardization, mini-model development and dissemination, and specialization.

THE MEANING OF ASSESSMENT

My first proposition--that assessment is and should be judged as an organic part of the use of models in policy evaluation--is not universally adhered to by the participants in this workshop. For example, more exalted criteria resting only on notions of scientific validity have been advanced.

I believe such a limited characterization of assessment is a poor guide to foresight or action because scientific validity is only part, and the relatively easy part, of the problem. To illustrate my case, let me consider the simplest serious instance of the validation problem.

Early in this century, the mathematician, David Hilbert, posed to the mathematical profession a list of questions, prominent in which was the problem of validating arithmetic--specifically, could the rules of arithmetic be used to prove their own internal consistency.

In 1931, Kurt Gödel proved you cannot validate the arithmetic--the consistency of the arithmetic is undecidable. This suggests that prospects for validating any models which use arithmetic are, in these terms, bleak.

This of course is somewhat fanciful, but the major contribution of Gödel's proof was to provide an absolutely explicit notion of the nature of a validation or the nature of a proof within the constructive and finitest rules that Hilbert's question had posed for the consistency problem. Gödel's result, which has been called the first theorem in social science, says something about what we are doing here. It says that within a validation esthetic which is that astringent, the demands for a satisfactory confirmation of the legitimacy of our efforts are to be frustrated. Now we are probably here for something that makes sense to all of us, so we can't mean validation or assessment in that narrow a construction.

My suggestion is that we are here to make more specific and to get on with an assessment process which occurs as a natural, organic part of modeling and the successful development of its use in economic policy assessment. If modeling is a part of the production of policy assessment, if it is viewed as an economic process which literally has had very rapid economic development, then the nature of what assessment ought to do now at this stage in the development of modeling becomes clear, specific, and palpable, and how it will evolve, if successful, also becomes clear, specific and palpable.

The punch line of this paper is that I believe validation and assessment activities are about communicating the model's representation of reality, and I mean that in ways which are very specific.

Hoff Stauffer presented what I thought was the perfect description of the modeler's agenda--to understand the reality which he is trying to represent and to evaluate his representation of this reality and its function within his model.

My paraphrase of Hoff's list of what the policy modeler must do is as follows:

1. Understand the phenomena which the model is to accommodate and represent in great detail.

2. Understand the issues which the model is intended to address, or, the questions the model is intended to answer.

3. Understand how these issues interact with the phenomena represented.

4. Ask if the model represents the relevant dynamics.

5. Ask if the model is structured to handle the phenomena and issues on the input side, in its internal structure and on the output side.

6. Ask if the data on which the model is based are well-documented and do they make sense.

7. Ask if the scenario assumptions are well-documented and do they make sense. Are they appropriate?

8. Is the model's output—the analytic results—intuitively acceptable? (If not, either model or intuition is wrong and one must be repaired.)

For what follows I will refer to this as the modeler's agenda.

I suggest then that formal validation and assessment activities should facilitate this process of assimilation and sensitivity testing, and should facilitate the understanding and use of its product, not just simply by the modeler and his peers, but by the customer as well.

In short, the validation and assessment process has specific clients—not just modelers, but the customers of the production process of policy evaluation. Formalizing validation and assessment activities and understanding their purpose will have specific impacts on this production process, and the success of the effort can be judged by whether or not the clients are finally satisfied.

So in designing and carrying out assessment activities, we must look to the reaction of the DOE's Professional Audit Review Team, the Congress and the policy community at large (as well as the scientific community) to see if these activities have been successful.

WHY ENERGY POLICY MODELING IS UNAVOIDABLE

Why has the Congress bestirred itself to make these demands for energy model assessment and to provide the resources and institutional framework for model assessments? I submit it is because energy modeling has turned out to be an unavoidable but extremely annoying adjunct to the policy process. Thus, since it cannot be ignored, it must be disciplined, socialized and assimilated.

Evidence that modeling is unavoidable in energy policy advocacy today is provided by the role played by PIES in the analysis and advocacy of the National Energy Plan.

Recall that PIES was originally developed by a Republican administration which was advocating an energy policy of price deregulation and accelerated energy supply development.

In the Carter Administration, the new Secretary of Energy is on record with, and is rather proud of, his extreme skepticism of large models. (See the lengthy and fascinating interview of Secretary Schlesinger in the Oil and Gas Journal, November 13, 1978.) Similarly, the Deputy Secretary, who is the cognizant senior manager of the Energy Information Administration as

well as the DOE's policy function, was a vocal opponent of the first proprietors of PIES in the 1975 debate over natural gas regulation.

Nevertheless, the PIES modeling group played a central role in integrating the energy market analysis of the Carter Administration's National Energy Plan (NEP). Within a month and a half after the new administration took office the PIES modeling group was contacted by members of the immediate staff of Secretary (then Energy Advisor) Schlesinger, and within a very short period of time we became part of an informally organized but very intense and focused analysis activity. Various groups developed assessments of specific measures, especially conservation and fuel management measures. The role of PIES was to provide integrated forecasts of energy market prices and quantity balances incorporating these import assessments. These balances and policy import assessments became the quantitative content of the policy dialogue between the Administration, the Congress and the public. For example, the energy balances on pages 95-96 of The National Energy Plan are from PIES solutions A148542C and A158569C.

An example of the impact of PIES and the necessity of a system like PIES in making the policy discussion intelligible is found in the evolution of NEP natural gas policy.

In the early version of the NEP, natural gas policy was motivated by the view that natural gas consumption was a moral problem. There were good uses of gas and bad uses of gas and the idea was to invoke regulatory prohibitions and immediate heavy tax penalties on the bad uses (mainly industrial and electric generation uses) of gas.

When these fuel management measures were analyzed in conjunction with conservation and supply provisions of the NEP, we found that about 6 or 7 Tcf of gas would be displaced from "bad" uses which could not immediately be reabsorbed in "good" uses. Since this gas had no market at prices competitive with oil, either the price must drop, inducing "non-economic" uses, or gas supply must be withheld, neither of which was a desirable outcome.

As the NEP evolved, its regulatory prohibitions and penalties on gas vis-a-vis oil were weakened, attenuated, and delayed until finally, in the bill that was passed, they weren't there at all. In the event the demise of the oil and gas user taxes has not been bemoaned by the DOE policy office. In fact, currently, the "gas supply bubble" is all the rage and the Secretary has effected regulations which are intended to induce the substitution of gas for oil in industrial boilers.

There are several lessons in this story. First, the most important contribution of a model like PIES is the accounting framework it imposes on the analysis. This forces the complex of policies to be specified with internal consistency, a property not usually present on the first try. It also requires that policy impacts be accounted for consistently, which makes it harder to double-count benefits and ignore costs.

Second, the initial version of the analysis described was performed by hand by the analysts associated with the model within 48 hours after receiving the policy description and analysis request. The formal model runs were completed several weeks later on a version of the policy which had already undergone substantial refinement. This shows (a) that you can't separate the use of the model from the use of the modelers, and (b) the really timely contribution of the model to policy development was

extremely informal. Formal model runs later served to discipline and validate this contribution and to bolster the subsequent advocacy process. If assessment procedure requirements stifle this kind of early informal contribution, they will debilitate the use of models in policy formation.

Among the microeconomic policy areas, energy policy may be unique to the extent that its discussion is carried on in quantitative terms. Thus, the discipline of a quantitative model turns out to be an indispensable adjunct to the orderly progress of the energy policy discussion.

Energy policy analysis can be distinguished from harder policy questions for which modeling "successes" haven't been nearly as prominent or notorious. Population policy, income distribution policy, economic development, believable environmental impact analysis are all areas where attempts have been made at modeling, but they haven't been nearly as successful, compelling, necessary or controversial. When I used to go around giving PIES talks people would ask me, "Is the energy problem so hard and mysterious that you had to build these big models?" and I would always say, "No, the energy problem is sufficiently easy and well understood that there is some point to building these big models."

There is essentially agreement or at least well established quantitative positions on the essential technical and behavior features of much of the energy system. Discussion and analysis of these features of the energy system in planning, in policy debate, and in regulatory proceedings are well developed. The discussion is going to continue to be carried out in quantitative terms. All parties will tacitly agree to a consistent quantitative organization of this discussion, and a big system model does this. The first presentation of a set of energy balances with and without the proposed policy inevitably becomes the framework and format for subsequent debate.

WHY ENERGY POLICY MODELING IS ANNOYING

That is why energy modeling is unavoidable. Now, why is it annoying? It is annoying because modeling itself is annoying, because science is annoying. Modeling essentially involves an abstraction, which, when you get down to talking about dollars and cents (or anything operational) is inherently questionable--always. That is the nature of science.

No one can use a literal geographical image of a country (whatever that would be) to plan a trip, and no one can use a literal socio-economic image of a country (whatever that would be) to plan a policy. But it is very hard to get energy policy-makers to understand and agree to the limited but useful legitimacy of a model's representation. Surprisingly, sometimes it is also very hard to get energy-policy modelers to agree to the legitimacy of a model's representations, and modelers have a vested interest in the perceived legitimacy of modeling.

Let me give an example. I am currently a member of the Demand Elasticity Working Group organized by the EPRI/Stanford Energy Modeling Forum. At the outset, the task of this group looked like a simple and technical one--namely, to clarify and record the prevailing wisdom about energy system demand elasticities. In fact what immediately emerged was a

fundamental confrontation over the legitimacy of the representation of the world as presented by process modelers versus the world as presented by econometricians.

Process modelers will specify a very explicit and responsible-looking representation of the process at issue, and then, essentially because of data inadequacies, employ ad hoc parameterizations of this process. On the other hand, the econometricians use ad hoc and mysterious ways of aggregating over the processes. They say things like, "Well, let's specify the industrial sector production function," whatever that describes. But then they have a very formal and rigorous way of extracting a parameterization of this construct from the data.

It emerged that the process models exhibit lower demand elasticities (on the order of -0.2 to -0.4) than those in the econometric models (on the order to -0.4 to -0.6).

Alan Manne effectively pointed out that the difference was consequential. If the elasticity of energy demand is in the range of -0.2, then the advanced energy technologies currently being contemplated are terrific deals and ought to be funded massively. If, however, the elasticity of energy demand is more like -0.5 then these projects are all losers and can only be justified on a contingency basis. With this higher price elasticity, people will be willing to forego expensive energy when the cheap stuff runs out, preferring instead to reduce consumption, reagglomerate production, distribution and commutation patterns and sit at home in the cold and dark (or, at least, the cool and murky).

Alan also reports that analytic experts who advocate a particular program will have a view of the nature of the energy system which is consistent with that program, even in areas where they have no special expertise. Environmentalist/conservationists believe that demand elasticities are large, resources are limited. and pessimism regarding technological/environmental capacities is warranted. High-technology advocates believe that demand elasticities are low, resources are abundant and technological/environmental optimism is accurate.

If the modeling professionals, who are trained to distinguish between form and substance, have this much trouble agreeing on the legitimacy of models, how much harder is it for policy makers?

There is evidence that the policy makers find appreciating a model's representation difficult. For example, at least from the modeler's point of view, they usually ask the wrong question in the heat of debate.

I can remember when an early PIES forecast of 1985 oil imports of three million barrels a day was confronted with an Exxon forecast in the neighborhood of 10 million barrels a day. The first question we were asked was who was right and who was wrong. Of course, the right question was--do the forecasts differ because of what we would call modeling differences or scenario differences. On consulting with Exxon we found we essentially shared the same "model" in our technical and behavioral characterization of the energy system. But the PIES scenario specification of immediate price deregulation and aggressive supply measures was designed to show what would happen if everything went right. The Exxon scenario specification

(in retrospect, an accurate forecast of the policy outcome) was for continued price regulation without aggressive supply policies.

Since the forecasts were conditional and were designed to answer different questions, neither was wrong, but this distinction between model and scenario specification and validity doesn't immediately occur to the policy-maker.

The policy maker has a model and a set of scenarios in his head which he brings to the policy analysis and advocacy problem. But for him the difference between model and scenario is less explicit and less articulated, just as the difference between analysis and advocacy is less explicit. But in compensation the policy-maker's model is much richer in the phenomena it relates because it hasn't had to be pruned into an explicit, internally consistent representation. Hence, since the answers lie in the assumptions, the policy-maker is not naturally sympathetic to the necessity of this pruning process before the analysis commences. (Note since it is not explicit, the policy-maker's model cannot be subjected to assessment.)

As a consequence, when the policy-maker inspects the assumptions made to get a tractable model, he tends to regard the model either as unresponsive because phenomena which might matter (you can't tell without analysis) have been excluded, or as hopelessly crude in representing phenomena of which he has a detailed and highly valued command. In such a case the policy-maker can perceive the modeler as an irresponsible but potentially dangerous charlatan in an area he feels he knows and owns.

To get the policy-maker past this kind of cultural response into understanding the limited but essential contribution of modeling is very difficult. It involves the almost metabolic process of sensitivity testing of phenomena, relationships, data issues and results, which is the content of Hoff Stauffer's modeler's agenda. That is, the client must go through a version of the assimilation and sensitivity testing process that the modeler went through.

Model assessment must facilitate this assimilation and sensitivity testing process. Therefore the requirements for reporting and procedure which assessment imposes on the modeling process must not only record compliance with established standards and procedures, they must explicitly evoke the outputs of the sensitivity testing of the model's representations in ways which make both model development and model assessment palpable, effective, and efficient. This is hard because responsible representation is hard. Thus if assessment in this sense succeeds at all, it succeeds at a difficult task.

Problems of understanding representations of reality are not specific to the energy modeling business. They have lain at the heart of scientific enterprise since it began. For example, if modern science was invented in the middle of the 17th century somewhere between Galileo and Newton, then it was marked at the start by an immediate conflict, ultimately a very consequential conflict, between the view of space as proposed by Newton and the view of space as proposed by Leibnitz, the resolution of which waited for Einstein. Kant made a tentative attempt to solve this problem when he said that the relationship between reality and our understanding of reality

has one invariant which is uniquely common to both--Euclidian geometry.
Kant said Euclidian geometry characterizes both reality and how we
understand it. That Kant's beguiling proposition turns out to be bad
mathematics and erroneous physics shows that this question of the validity
of reality's representation is not simply philosophic but is consequential,
is generic to science, and is difficult to resolve.

We should also understand that model assessment isn't unique to energy
modeling. There was a period in history when questions about the number of
planets and the shape of their orbits were very prominent policy questions,
and the assessment of the analysis of these questions wasn't carried out by
the Office of Analysis Oversight. It was carried out by the Office of the
Holy Inquisition.

WHY ASSESSMENT NOW?

Why has the demand for formal model assessment arisen now? I think it
is because communication about the nature of the ingredients and the
functions of modeling is extremely important and difficult at precisely
this stage in the economic development of modeling within the policy
analysis process.

I view policy modeling as an indirect means of production of policy
evaluation in precisely the sense that David Ricardo meant the phrase
"indirect means of production," that is, the production of commodities by
means of commodities.

When a farmer plows his field by hand, he is engaging in a direct act
of production. When he, or more generally, a society of which he is a
part, spends part of its time and effort building a tractor to plow the
field, that is an indirect means of production. In policy areas which are
harder, more complicated, and less formalizable than energy policy, the
acts of policy analysis and policy advocacy occur within the same head, and
there is no separation of the analysis machinery from the advocacy process.
We see such separation when model development becomes a separate and
explicit activity, an intermediate output of productive activity in the
production of policy analysis.

Thus modeling is, in this metaphor, capital intensive and indirect.
What are the consequences of this? We know them from Adam Smith and from
the economic development literature. Economic development elaborates
intermediate products, generates infrastructure investment, and induces the
specialization of effort. This is what will be happening to modeling in
very specific ways, as I will suggest.

For this development to happen, markets in which the intermediate
products are understood and traded have to be established. The assessment
process must provide the social infrastructure which supports those
markets; that is, it must provide the education, the communication
networks, the transportation networks, the kinds of standardization of
taxonomy, standardization of product, validation of product quality, and
other informational externalities that facilitate the market for the
intermediate products--models and their results--which finally produce
policy assessment.

This is neither trivial nor obvious. One could write down a much less ambitious program--and some of the people who spoke here yesterday have advocated such--for the assessment process. Perhaps model assessment could be content simply with scientific peer review rather than with trying to make sure that Senator Jackson's staff understood it.

We can try that on for size. We can suppose that George Lady as the Director of Analysis Oversight goes over to Senator Jackson's office after an appropriate ceremony and says, "Senator Jackson, we have subjected this econometric model to peer review, and you will be pleased to know that a collection of wise and profound econometricians has said this is a terrific model because it has been estimated with the iterative Zellner technique which everybody knows is asymptotically equivalent to full information maximum likelihood estimation." You can just imagine Senator Jackson saying, "Whew! Boy, that satisfies me, but I was really worried there for a minute." Right?

On the other hand, Hoff Stauffer suggested that we might evaluate models in terms of their answers. I think that is wrong, too. Nobody in the room is going to be satisfied with the statement that such-and-such is a terrific model because it says we need to put the "cost of work in progress" in electric utilities' rate base or we need to deregulate gas, or whatever.

Assessment cannot be satisfied simply with reporting a scientific consensus on legitimacy of procedure. Nor can it be content with evaluating model's answers. Assessment is being invoked to communicate the nature of models and their ingredients to the policy-making community.

IMPACT ON THE MODELING PROCESS

What will be the impact of assessment on the modeling process? First, there will be increased emphasis on modularization--the identification of models and of sectors of models and their associated data bases as separate products which themselves need validation and assessment. Emphasis on modularization will make possible all of the other dimensions of specialization in modeling activities.

Second there will be the identification of what I call high level data as products in their own right, and this is important institutionally both for the people who produce these kinds of data and the people who use them. By high level data I mean numbers which energy system modelers regard as data but which sector modeling or sector analysis efforts regard as output: things like "finding rates" for resource exploration, load duration curve representations for electricity demand, capital and operating costs, capital charge factors and the like. These kinds of numbers are inputs--data--to an energy systems modeler, but they are the final product of a complicated analysis process to a person whose office says oil or gas or electricity or coal on its door.

The fact that these numbers are complicated analysis products doesn't mean they can't be regarded as standardized products. For example, the Consumer Price Index proposes to measure a very subtle and profound

phenomenon, but nevertheless, for most purposes, the CPI calculation is regarded as data produced by a meaningful, reliable, standardized computation used by both sides in policy debate over, for example, the minimum wage level or income tax policy.

Research chemists no longer make their own sulphuric acid. Similarly, systems modelers will draw away from de novo data generation. The kind of data generation process which Hoff described in building the data base for the National Coal Model will be specialized, institutionalized and standardized. There will be, therefore, institutional recognition of the products and of the institutions that produce these sector analyses inputs. This will be accomplished through a sorely needed revolution in the Energy Information Administration wherein <u>data for analysis</u> will achieve coequal status with data for regulation as a management and budgetary objective.

There will be <u>increased specialization</u> of energy sector analysis and energy systems analysis operations. Distinct organizations can organize different kinds of expertise and different standards of successful performance will be elaborated.

There will be a strong thrust in developing <u>mini-models</u> of sector models, both because they communicate a representation of a bigger model and can be easily compared, and because they can be used interchangeably in assembling systems models. Documentation of big sector models should include standardized mini-model representations of these models to summarize model behavior.

Competition and comparison between mini-models will induce <u>standardization</u> of sector definitions, commodity definitions, technology characterizations and sector interface gates to facilitate comparable measurements of value and commodity flows in model solutions.

This will facilitate the comparing, interchanging, and reproducing of models and model components, so that these become reuseable "capital goods." It also means that the documentation, validation and assessment of model components can be efficiently entrained in the documentation, validation and assessment of energy system models of which they are constituents as they are assembled and reassembled.

There will be development of <u>systems software</u> in which model specification syntax, model solution algorithms, data structures, data handling and report writing all exploit the common structure of the model. (The modeling system developed by Cazalet and his associates in which DOE'S LEAP model is written has this integrated structure.)

Finally there will be <u>specialized research</u> in mini-model representation--how to write little models which represent big models in some satisfactory sense. Similarly there will be specialized research in systems models--in solution-existence, in inter-sectoral interactions such as regulated behavior, monopolistic behavior and other deviations from competitive optimizing specifications, as well as algorithmic research exploiting mini-model and systems model structure.

WHAT TO DO?

What should we do to facilitate the symbiotic growth of models and model assessment in policy analysis? First and most importantly, I think the focus on modularization and standardization is crucial. It is the only way to break out of the giantism and insularity (the not-invented-here syndrome) that afflicts all large-scale modeling efforts. It is also the only way efficiently to admit reproducability into model performance and model analysis, and this is crucial if the mystery surrounding models is to be dispelled.

Modularization makes for better modeling too. One of the strengths of the PIES operation was that the output of many different analytic offices had to be communicated to the integrating framework program in explicit data files. This facilitated the specialization of labor, focusing substantive expertise. It also fixed responsibility for sector model performance, and it surfaced anomalies efficiently. Because the institutional organization forced each group to look at other groups' formal outputs to diagnose malfunctions, we often found we had a much deeper understanding of system modeling and its pitfalls than did modeling efforts in which the entire analytic process occurred within one piece of code.

Second, hand in hand with modularization, modelers--both system modelers and sector modelers--should publish mini-models of their models (Taylor series expansions of model behavior). It should become standard practice to publish, in standardized form with standardized definitions and measurement conventions, the mini-model expansion of a model's behavior with respect to the important internal parameters and with respect to the scenario variables.

Third, there ought to be a ruthless institutionalized focus on the data in the models. One way to do this is to force modelers to publish the output and input of their models in the formats in which real-world data systems publish these data. This should be true for the model as a whole and for the sectors in the model, with a lot of attention to the sectors.

This will have three impacts: First, it will mean that the modelers have to understand how the world measures these things and what they are, which currently is not always the case. Second, once the model has produced its version of these data, the user of the model is looking at something with which he is, in principle, familiar, so that his expertise is immediately challenged. Third, it will focus on what the real-world data systems are actually doing, and I suggest that, in general, we will find that they don't represent what they say they do, and they don't measure accurately what they do measure.

Another way to force a focus on data is to require the data people to attempt to recast their data in the way their modelers use it. I propose this as an immediate and important exercise.

The EIA's Monthly Energy Review currently really comes from the old Bureau of Mines data systems. It shows its phylogeny from the Bureau of Mines' view of the world where the Bureau had an operating mission, not a data mission. Thus, the MER amounts to a kind of single entry bookkeeping representation of what important energy entities have been doing this

month, but it doesn't have any of the internal accounting structure by sector that disciplines a comparison of the balances beween supply and demand which is inherent in a modeling structure like PIES, or the LEAP Model, or the Brookhaven Energy Reference System. (The MER does notionally allocate consumption into the original BOM fuel-sector consumption breakout developed by Walter duPree and James West but this accounting structure is not used to structure or discipline data gathering.)

It would be a fascinating and timely exercise if the data published in the MER and the other data collected by EIA were cast in terms of the network or energy reference system structure which the energy modelers have been working with for at least 10 years. This is difficult to do. Ask Ken Hoffman, who is the father of the Brookhaven modeling effort, how hard it was to create the 1972 reference energy system, and he might say "Well, it was really hard, and we are still not sure that we got it right."

Finally, there is an effort which hasn't really found a use inside EIA, and that is the National Energy Accounts, which are cast in the accounting conventions of the national income accounts. If that accounting system is used by the modeling community, and there is an effort to do so under contract to DOE by Dale Jorgenson and his associates, then the accounting for value flows and factor flows and a general view of how energy interacts with other factor choices will be taken farther forward. This is absolutely crucial in providing a scientific basis for assessing conservation possibilities and programs or for understanding energy-economy interactions.

As I have mentioned before, all of this data development awaits recognition that data for analysis must become an institutionalized product with high priority for the top management of the EIA. Energy data now is in the same state as national income data was when it was in the hands of the Customs Bureau and the tax agencies. It lacks an accounting discipline, a statistical sampling discipline and an analytic measurement discipline. No amount of bootlegging regulatory data bases will provide a scientifically valid basis for energy information.

Here really is the most important target for energy modeling assessment. If the data systems are made valid, valid science will surely follow. Most specious ingenuity in modeling arises to bridge data gaps.

DANGERS OF ASSESSMENT

I want to comment very briefly on the downside of the assessment process. What are the costs of institutionalizing assessment requirements? First, assessment is very expensive. One modeler I was talking to at this workshop said that when he was approached about cooperating in a third party assessment activity, he suggested that they give him 10 percent of the proposed funding and he would do a better assessment, and if they gave him the other 90 percent of the money, he would fix the problems it surfaced. That may be a bit flamboyant, but it is true that assessment costs a lot of money.

On the other hand, any infrastructure investment costs a lot of money when you divide it by the output of the infant industry you are trying to facilitate. If it is a growth industry, then the unit costs will decline.

Assessment might make modeling less immediately responsive. This isn't necessarily a bad thing. I often felt when I was working for Bill Hogan in the early days of PIES as though I were a cab driver going down Pennsylvania Avenue and Bill would get in and say, "I want to get to Dulles Airport in four minutes." I would say, "Well, that is a very dangerous thing to attempt." He would say, "Yep, but if we don't get there, the plane leaves." Bill's clients were in a hurry. If they didn't use our numbers, they would use somebody else's. We probably had better numbers than somebody else, so it wasn't responsive in that environment to say, "I can't give you an answer."

Further, all modelers are subjected to false deadlines by people who don't understand the realities of modeling, and sometimes in response to this you sell loss leaders, so you try to fulfill these demands. You adopt a strategy of maximizing the probability of getting the answer right by tonight, even though that is a dangerous strategy because when you don't, you have to start over.

Institutionalizing the requirements for assessment might make for realistic deadlines for analysis inputs to the policy discussion, and it would make modeling a very much safer activity. On the other hand, it could also slow down things so that modeling "in crash mode," in Harvey Greenberg's phrase, would become infeasible. This would be too bad, because policy analysis, when it really matters, usually occurs in crash mode.

What we have to be careful about is letting assessment standards be flexible enough so that different findings assembled under different time constraints have available different colored covers which represent different degrees of validation and assessment. Modularization that entrains the long-term validation of the ingredients of a model and its findings really contributes to this differential validation requirement.

Finally, there is, I think, a danger of totalitarianism. I am not sure that on the whole the regulation of modeling is any better for modeling than the regulation of oil and gas production is good for oil and gas production, and I think we have to be careful about how much power we give to the assessors. We ought to remember that after the Inquisition finished its assessment of Galileo's work there were no telescopes built in Catholic Europe for centuries.

Thank you.

DISCUSSION

Dr. Mayer (Princeton University): One of the issues that you brought up in the several times that I have heard you speak is this distinction which I think is false between people who use models in policy analysis and people who are kind of mush heads. You say, "Well, they have a model in their own brain and that you can't validate because they either use a model or they don't." I am thinking in particular, and I would like for you to address, for example, biomedical research, where we do things like

estimate the number of people who are going to die of particular cancers due to cigarette smoking, or for women, due to taking the pill.

In fact, these judgments and forecasts, which are probably more accurate than anything we have in the energy area, are made by a very sophisticated combination of what I would call judgment and analysis, and where appropriate, that analysis is a model such as basic biochemical models of reactions of the human body to particular drugs.

The question I have is: How do you get from the idea, or what makes you think that if someone, let's say Jackson's Committee, is not going to use a very sophisticated 70, 80, or 950 equation model that that doesn't mean they are not using any analytic analysis? Somehow you seem to equate a large-scale model with being analytic. I don't quite understand how the two are equal.

Dr. Nissen: I don't think I said that. What I said was that there are areas in which analysis is carried out without formal models that have computer printout. I think, for example, all of the policy that is made about income distribution in the United States has that characteristic. A lot of the policy that is made about regional environmental impacts is made on that basis, although we are making very slow and very painful progress.

But I was just distinguishing easy problems and hard problems. I am not against implicit models. I am suggesting that there is an economic process and rationalization to investing in a distinguishable modeling product.

Dr. Mayer: But I am asking you just--I want you to go on the record. Could one, though, be totally analytic and use models, and use a totally analytic framework, not anything implicit, without using a large-scale model?

Dr. Nissen: Sure.

Dr. Hyde (University of Maryland): There is sort of a question that was raised in my mind from yesterday, and you sort of touched on today, and that is the place of peer group assessment. I still seems to me it might be the most cost effective way of analyzing these models and seeing which ones really are the best. I was curious. Were you attacking that or possibly supporting that idea?

Dr. Nissen: I am saying it is a necessary ingredient. Scoop Jackson's staff is going to go to other people and they are going to say within the accepted conventions of scientific methodology it is okay; people ought to say, yes, and it ought to get looked at. But I am saying that is not enough.

An example I gave demonstrated that it wasn't enough. It is not going to amount to effective model communication to the staff. It is an ingredient in it, but it isn't the whole thing.

THE ENERGY MODELING FORUM AND MODEL ASSESSMENT: SUBSTITUTES OR COMPLEMENTS?

John P. Weyant
Department of Operations Research
Stanford University

INTRODUCTION

I have chosen quite a provocative title, I think, "The Energy Modeling Forum and Model Assessments: Substitutes or Complements?" First I have a confession to make. As is often the case with such titles, it is more provocative than it is an accurate summary description of what I am going to say. Indeed, I could pose the problem that I am going to address as to maximize the credibility and acceptability of energy policy models by adjusting the relative funding levels to model assessment and EMF-like activities within a fixed budget constraint. That I will not do.

What I will do is to take a somewhat softer tack and talk in more general terms about the relative strengths and weaknesses and what I feel to be the appropriate roles of the two types of model analysis activities. In that regard, a more accurate title for my talk might be one that David Wood suggested to me recently, "The Energy Modeling Forum and Model Assessment: Where Do We Draw the Line?"

I would like to say a little bit about my own personal perspective on this subject. First, it has been my privilege to have participated in the Energy Modeling Forum (EMF) activity from its inception, thanks to Bill Hogan and Jim Sweeney. Second, I have always had what I considered to be an implicit, obvious interest in model assessment activities and projects solely because I have felt that they were of great relevance to what we do at the EMF. In writing this paper, I have forced myself to make that relationship a bit more explicit.

The first thing that I can think of in the literature chronologically that is relevant to this topic is the book by Greenberger, et al. entitled, "Models in the Policy Process" [1]. In Chapter 10 of that book, "Modeling and the Political Process," which is their conclusions and recommendations chapter, they point out that there indeed seems to be a problem with the actual use of policy models in general, and that there are lots of things that we can do to improve this situation. We could create better public education in modeling techniques; we could create more professional standards amongst the modelers; we could create more responsibility on the part of the decision makers in commissioning model studies; we could better define model assessment procedures and documentation standards; and finally, we could create bridges between energy model builders and model users.

They astutely observe at that point, however, that if all of these courses of action were pursued simultaneously, none would get done very effectively. So they conclude that the model assessment activity is probably the highest payoff single activity. Indeed they imply, and I will say explicitly, that there is a lot of overlap between that particular course of action and the other four.

The next piece of relevant literature is the paper Bill Hogan presented at the Lawrence Livermore conference on October 3rd of last year: "Energy Modeling: Building Understanding for Better Use" [2]. What I would like to recall from that talk are the three types of model evaluation activities he described: verification and validation -- for which

we are converging on standard definitions [3] -- and one Bill added to that, ventilation -- simply opening up the architecture and structure of the models to public scrutiny. We were also reminded in that talk that the purpose of modeling and analysis is not numbers, but insight.

The third thing I would like to draw on is a lecture that David Wood gave at Stanford on November 9th of last year, entitled: "Model Assessments in the Policy Research Process" [4]; That lecture provoked both the title and substance of my comments. Finally, I have tried to draw on and in some cases integrate the comments and observations that I noted during the presentation of the other papers at the conference. Indeed, it turned out, in going through my notes, that almost everyone had something to say that was relevant to my topic.

In the next section, I will quickly review the ongoing model assessment and EMF-like activities. Then I will recall a set of requisites for energy policy models that I will use to talk about the relative strengths and weaknesses of the different types of model evaluation and model analysis activities as currently practiced. Then I will move from that rather abstract discussion to a more operational accounting of what model assessment activities currently do that the EMF does not do, and conversely, what the EMF does that the model assessment activities don't do today. That leads quite naturally to an admission on my part that the dichotomy that I have created between assessment activities and the EMF is actually not a very precise one after all; there is a fair degree of overlap. There are many aspects of the EMF process, as now configured, that are really just different types of model assessment. Finally, I will talk a little bit about some alternative directions for the future for these two types of activities.

ONGOING EMF AND MODEL ASSESSMENT ACTIVITIES

I don't want to spend a lot of time reviewing the ongoing model assessment and EMF-like activities; especially since most of the other papers presented at the conference are focused on that task. Additionally, Greenberger [5] provides an insightful way of thinking about these two approaches to "model analysis."

I am interested here only in model assessment (MA) activities that go all the way to the hands-on stage. There are several of those that we have heard about. The first one that comes to mind is the MIT Model Assessment Laboratory [6, 7, 8, 9]; the second is some of the work that they are doing on the Texas National Energy Modeling Project [10]; and the third is the model assessment work going on in DOE at the EIA [11].

The EMF-like activities, which I define as those not only doing standardized model comparisons, but also involving users in that comparison process including the Energy Modeling Forum [12], the Utility Modeling Forum [13], a serious desire at the Solar Energy Research Institute (SERI) to do a solar energy modeling forum, and an exercise carried out by the EIA in the preparation of the 1978 Administrator's Annual Report.

USING MA AND EMF ACTIVITIES TO IDENTIFY ACCEPTABLE POLICY MODELS

In his Stanford lecture David Wood first defined a policy model as one that can be used to address an issue where there are a lot of contention points, as opposed to a research model where we are just trying to learn about a system per se [4]. He then identified a short list of requisites for a policy model. I would like to use David's requisites as a framework for talking about what I feel are the comparative advantages of the three different types of model evaluation activities; verification, validation, and ventilation. Then, I will overlay on that my assessment of which of the two generic activities, that is EMF-like activities and model assessment activities, are most effective in accomplishing the three model evaluation functions.

The first requisite for a good or acceptable policy model that Dave mentioned was that it must be based on good research results; that is, tried and tested principles, theory, and data. Here I think we are talking about what people have generically talked about as verification and validation [3]. Further, I think there was a hypothesis postulated by Dave Kresge yesterday [6] that the only was to do these things right is to do hands-on model assessment. Basically I agree with that, so I see in this first requisite that the way to determine whether an energy policy model is based on good research results is to do an in-depth model assessment. By that I mean hands-on, third-party independent review.

The second thing that David thought was a requisite for a policy model is that it should include all the relevant policy options; It should have the correct slope, consider the things that are important in the policy debate, and include the relevant policy levers. To be able to deal with the policies that are actually being considered has been a problem in the past because many of the models really don't have the right policy levers. So when a debate comes up, which is usually quickly, it gets resolved very quickly too, and the model is not of much use. The model should also account for the important impacts of the alternative policies. The third requisite, which I would like to lump with the second, is that the model should include the appropriate inputs and outputs to enable one to analyze the alternative choices, the contention points under consideration.

I personally feel that the kinds of activities that will best demonstrate and illuminate how the models stack up on the second and third requisites and help people evaluate the models in those regards, are basically ventilation-type activities. I think that is where the comparative advantage for ventilation activities are at this point in time. I perceive that there is a communication gap that exists right now in these areas that is important to mend before digging into the level of detail that one would go to in an in-depth, hands-on, third-party assessment. We should talk frequently in public about the policy levers and calculated impacts included in the various systems. This is why I believe that EMF-like activities have a comparative advantage in terms of increasing model acceptability and use through ventilation according to the second and third requisites at the present time.

RELATIVE ROLES OF EMF AND MA ACTIVITIES: CURRENT PRACTICE

In discussing the relative roles of EMF and model assessment activities some aspects of the current practice are useful reference points. In particular, it is important to identify the kinds of things that the model assessment activities do that the EMF does not, and vice versa.

What do the model assessment activities do that the EMF doesn't? First, the EMF does not do comprehensive "overview assessments" like, say, the Model Assessment Laboratory [6]. We have done a pretty good job of telling the model users what the models as a set in each study are good for and, importantly, what they are not good for. But, we have not done enough about explaining to the model users what the <u>relative</u> strengths and weaknesses of the individual models are on a comparative basis, even strictly in terms of scope -- what is assumed exogenously and what is left endogenous to the model. And maybe we could add a little bit about how the endogenous part gets done without doing any validation.

The other thing that the Assessment Lab does that the Forum doesn't do is hands-on, third-party independent assessments. That policy was recommended at the workshop that help set up the Forum, and the model assessment project as well, in the summer of 1976 [14].

What does the EMF do that the model assessment projects don't at present? Number one, it involves model users directly in the modeling process, and that is a good way, I think, to accomplish ventilation. The forum process, which amounts to a very focused encounter group type of experience, accommodates the acculturation to models very expeditiously. And an important product of the Forum, and certainly a product in which the Forum has a comparative advantage over the MA activities, is alumni, an increasing community of, we hope, born again modelers and born again users.

Another thing EMF does do that MA does not is standardized model comparisons. As I think we could gather from the several overviews of the Model Assessment Laboratory yesterday [6, 7, 8, 9], there are many people who feel that the assessment projects should do standardized model comparisons. Now, as Marty Baughman pointed out yesterday [8], there is in the present scheme an overview type comparison where the model that is being assessed is compared with published results from other similar models, and so forth. But, what I have in mind by comparative model assessments here is to actually run the same tests on more than one model.

The benefit from doing this type of comparison is that you can identify differences in the explicit assumptions made a priori and the implications of those different assumptions ex post. The example that springs to my mind involves what is probably the model now most widely used by the policy and evaluation part of the Department of Energy [15]. During the course of the second EMF study, the representatives of that model observed that the projection of percentage electricity generation by coal from their model was different than everybody else's. They then

discovered upon reconsideration that their model was not calibrated cor-
rectly. In other words, they had the wrong number for 1975. This type
of finding has come out quite dramatically in some of the model compari-
sons that we have done thus far.

The second benefit from doing standardized model comparisons, that
Jim Sweeney pointed out yesterday [12] and that is often overlooked, is
that by doing comparisons between models of different generic types one
can compare and contrast the differences in the impacts due to their dif-
ferent implicit assumptions. Dave Nissen [16] gave me a very good lead
in here by talking about the utility of examining the differences between
projections from econometric models, process models, and simulation models.
I think EMF-4, the energy demand elasticity experiment [17], is a case
where we will get deeply into that type of comparison.

Just what does a difference in world view mean in terms of the results
of a policy model? I would like to recall again an example that occurred
in the second EMF study, "Coal in Transition." We actually found a coal
model where the implicit implication of the objective function was that
a perfectly price discriminating monopsonist represents the behavior of
the aggregate consumer of coal in the United States. That might be accurate,
but the people who were doing the work at that point thought that they
had the paradigm of the perfectly competitive model imbedded in their
model's structure. They were surprised to find out what the objective
function they were using really implied. It may be overstating it to say
that they have since modified their model, but they now provide the option
of using either one of the two objective functions, and this is reflected
in the final report [18].

LEVELS OF ASSESSMENT IN CURRENT MA AND EMF ACTIVITIES

It should be evident from my comments so far, though, that the dichotomy
between EMF and MA activities has been weakened and that there are aspects
of model assessment included in the EMF activities as defined here. Table 1
is my reconstruction of the Model Assessment Laboratory's schematic diagram
of the different approaches to model assessment that you heard about at
some length yesterday [6].

The four different approaches seem to be: literature review, overview
assessment, independent audit, and hands-on assessment. I also listed the
key components of each approach. My little checks, "√", here correspond
to things that I think we already do in the Forum. Those of you who have
participated in our studies have been consumers of our reports and can
judge for yourselves how well we have done these activities. The components
I've marked with X's, on the other hand, are the ones that I think we should
be doing but haven't done very well so far. I must, of course, take much
of the blame for that. We actually do some of the independent audit function
focused on a particular policy issue. I don't mean to say that the Forum
has done this in the depth that the Model Assessment Project has where they
can run many mini-scenarios because they have only got one model. But the

TABLE 1

LEVELS OF MODEL ASSESSMENT IN MA & EMF ACTIVITIES

Literature Review

* Objective of Model √
* Appropriateness of Structure √
* Plausibility of Results √

"Overview" Assessments

* Model Logic
* Empirical Implementation
* Comparative Evaluation X
* Documentation
* Structural Limitations X
* Contention Points X

Independent Audit

* Experimental Analysis
    - Sensitivity √
    - Test Data
* Policy Impact Analysis √
* Documentation
* Contention Points √

"Hands-On" Assessment

* Experimental Analysis
    - Sensitivity
    - Test Data
    - Alternative Structure
* Policy Impact Analysis
* Replication of Results

operation of the Forum so far has meant that to find out about the models, we request that the modelers make certain scenario runs to find out what their output are. So the policy impact analysis is something we do in a comparative mode, much like the independent audit. We do some sensitivity analysis, and I think we do a little bit on contention points, but not as much as we should. We have also done a reasonable job in the past in reviewing the relevant literature.

The thing that I think we haven't done that we should is a more telling comparative evaluation of the relative structural limitations of the different participating models and their relative strengths and weaknesses in analyzing the relevant contention points. I think we have done a lot on the limitations and strengths of the set of participating models, but not enough about the relative strengths and weaknesses of the specific models participating in the studies.

RELATIVE ROLES OF EMF AND MA ACTIVITIES: DIRECTIONS FOR THE FUTURE?

What types of modeling and model analysis activities should we do? What should be the relative emphasis on each? Let me first go to some extreme point recommendations. Before we do anything I think we should consider what some extreme possibilities are. One is that we could stop modeling and then we wouldn't feel obligated to do either the EMF-like or model assessment activities. I think that is basically a cop-out. We would definitely lose the insights to be gained from formal analysis by doing that. Another thing we could do is just stop evaluating and assessing. That would ignore the main conclusion of Greenberger, et al. [1]; that model assessment and evaluation is a promising solution to the problem of lack of use and implementation of policy models.

Another extreme point recommendation would be to stop modeling until we know how to do assessment. I think that would again ignore the value of the insights provided by the models. I think Bill Hogan was very pragmatic in his Lawrence Livermore talk [2] in his recommendations about what to do about counterintuitive results from a model; Assume they are wrong for the time being, but at the same time try to figure out why the counterintuitive result might be correct. Then when you figure out why, it is usually either because there is a mistake in the model, in which case you might say, well, we told you the right thing to do anyway, or an insight occurs and at that point you can adjust your policy recommendation. An argument against this strategy is an argument against models ever providing insight; something that is inconsistent with the evidence.

Based on my observations about the overlap between the two kinds of activities, another thing one could do is to combine the model assessment and EMF activities into a single activity. One could design a forum study and at the same time simultaneously do an in-depth assessment of each model; We could just bring all of the models up at Stanford. The problem with that course of action operationally, I think, is strictly resources.

Jim Sweeney was probing Dave Wood last night at dinner, and I think Dave agreed with Jim that to assess one model costs about the same as it costs to do a whole forum study. Additionally, the production of alumni, I think, would be materially impeded by getting much more ornate about the EMF operating the models on its own. I think that the fact that the Forum cannot operate the models forces meaningful dialogue to take place between modelers, and between model builders and model users. This is a substrate of a very important product which the Forum is producing, an increasing community of well-educated model builders and model users.

In the initial stages of the two types of activities, I had always implicitly conceived of the EMF as being a screening device for the Model Assessment Project. I was interested to hear Dave Nissen ask Bud Cherry the question the other way around yesterday. If one or more of the models in the Forum load-forecasting study [13] had been through an assessment, wouldn't that give you a lot more guidance and resolve a lot of the issues that came up in your study? To my mind, it is really not clear which should drive which. But the recommendation I would make here is simply that we should better integrate the two activities.

Finally, I am going to conclude that I think the directions that the two activities are taking now are the optimal ones. First, I think that it is probably a good idea for the model assessment-type projects to do comparative assessments. I think they are doing some of those in the Texas National Energy Modeling Project [10] and they have thought seriously about doing them at the Model Assessment Lab. In the EMF, at least the way things are presently done, you really lose part of your control over the comparison because you rely on the modeler to go off, do his runs, and so on, and you develop some insight, but not a very in-depth insight into just exactly how a modeler implements a particular policy. In a more controlled situation, which we would have were the comparisons done in MA, you would get a lot deeper into the comparison, and gain a lot more in-depth explanation of what different structural and data elements in the model account for differences in model result. There will be an increase in resources required to do this, however.

So to answer my initial question, I think the model assessment activities probably need more incremental funding now than the forum-type activities. But, additionally, I think the EMF has a strong comparative advantage to do certain types of assessments, and I think I have made those clear: just a more telling comparison amongst the various models as to the policy levers in them, the relative impacts that they report or could calculate in a meaningful way, and how well they can address various contention points. This too will require some additional resources. There is a natural complementarity between the two types of model analysis activities that could be more fully exploited with a little forethought and cross fertilization. Both activities can and should benefit from the experiences of the other.

Dr. Nissen: Your talk concentrated an awful lot on model assessment as an output of the forum. It occurred to me, and I would like to attribute this insight to Dave Wood, who dropped a remark like this, that at least an equally important product of the forum, and certainly a product in which the forum has a comparative advantage, is alumni.

The forum workshop process, which amounts to a very, very focused encounter group, accommodates very expeditiously the acculturation to models that I was taking about as a product of the assessment process. The production of alumni, I think, would be materially impeded by getting much more ornate about operating your models on your own.

I think that the fact that the forum cannot operate its models and therefore communication between modelers, and modelers and model users, is a substrate of a very important product which the forum is producing, and that is an increasing community of, we hope, born-again modelers and born-again users.

Dr. Weyant: That is a more formal statement. Just to increase ventilation, I think, would proceed that way.

Dr. Wood: Well, I would like to make a couple of comments. I am still waiting for John to tell us whether we have substitutes or complements here.

Dr. Weyant: To ventilate; the number again in Palo Alto is 405 --

Dr. Wood: The number is MIT--I guess what I distilled from John's comments, though, is that we have complementarity and that there is clearly quite a bit of overlap between the objectives and the activities of model assessment and the objectives and activities of the MF.

One comment I want to make is I can't believe that I said, "Involving model users in the modeling process is second order to model assessment."

Dr. Weyant: That was in terms of increasing model credibility.

Dr. Wood: Yes.

Dr. Weyant: You did qualify that by saying that if one wanted to organize a modeling research project, that involving users was right on.

Dr. Wood: Right, right. Exactly. But I see sort of different objectives being served there. Clearly, I think one of the biggest problems with the model development process is that we haven't found ways to involve users sufficiently or to get resources to involve users sufficiently. That is one of the problems that ventilation it seems to me is retrospectively trying to address.

I guess I will make a comment about what kind of comparisons we might expect between the EMF, as it is constituted now, and model assessment (MA) as it is constituted now. I think they would be very different, and I think it would be misleading to think that they are in any way comparable.

I think in EMF you will, at least the way things are presently done, you really lose control because you rely on the modeler to go off, do his runs, and so on, and you develop some insight, but not very indepth insight into the problem that David Kresge was talking about yesterday; namely, just exactly how does a modeler implement a particular policy?

In a more controlled situation, which, if we were doing these things in MA we would have, I suspect you would get a lot more indepth into the comparison, a lot more indepth explanation of what different structural and data elements in the model account for differences in model result. I am reminded of what I always think of as kind of a classic little piece that was published in the "Bell Journal" a couple of years ago, where it sorted out the differences between supposedly comparable forecasts of the MacAvoy/Pindyck Model and the AGA Tera Model, I believe it was.

It is a marvelous little case study in what one has to do in order to track what parts of the model and what data elements actually account for differences in forecasts. That is the kind of comparison you should get out of model assessment. It is the kind of comparison that seems to be difficult to get out of the EMF.

I think this comment about the cost of model assessment is an interesting one. You are probably right when we are talking about a full indepth kind of assessment. Model assessment is much less expensive when one is pursuing it to the level of, say, what David Kresge--well, I really refer to David Kresge's conversation yesterday about independent audit. The cost of getting through an overview and an independent audit are fairly low compared to that last step.

That leads me to another observation. If we internalize model assessment in the development process so that we are satisfied that there has been a kind of independent assessment at an appropriate level during the relevant stages of model development it will introduce considerable efficiency in the assessment process. The aggregate integrating over all those costs will be much less.

Part of the large expense in assessment, at least the ones that we have done, has been that it is a starting at square zero sort of thing. If we had been involved in the earlier stages, we would have been much more efficient in the process. I think Rich Richels made a point yesterday about follow-on audits, and so on, where they would be much more efficient and cost effective.

Well, there is a symmetric part of that. You can extend that in the other direction. I guess I will make one last comment. It seems to me that you agree with me that much of what we are talking about here is the development of sound practice into the realm of policy model development and application. As we invent some organizational initiatives such as forums, as we invent some guidelines for assessment, as we educate sponsors of model development to ensure that their contracts and RFPs generate the set of materials necessary for assessments and independent observation on a model, we will find what seem to be differences between a forum activity and an assessment activity are really going to begin to disappear. The lines between them will become very blurred, and the industry of model assessment will have a relatively short life.

It is an industry that is worth developing now because it focuses attention on some serious problems that are inhibiting the credibility and use of policy models, but it will probably serve its purpose fairly rapidly if we are successful and we won't in a few years talk about forums versus model assessment. We will talk about modelers and model users in the policy research process, and we will get the kind of interindustry activity fairly well established that Dave Nissen was referring to.

We are all in the business of producing policy evaluation. We are trying to invent right now some intermediate, some industrial processes, if you will, that are intermediate processes and we are highlighting them in the process; but in the long run we will probably cease to highlight them.

Dr. Weyant: I have a lengthy response to those comments, but I have already taken more than my fair share of time.

References

[1] Greenberger, M., M. A. Crenson, and B. L. Crissey, <u>Models in the Policy Process</u>, Russell Sage Foundation, New York, 1976.

[2] Hogan, W. W., "Energy Modeling: Building Understanding for Better Use," paper presented at the 2nd Lawrence Symposium on the Systems and Decision Sciences, Berkeley, California, October 3, 1978.

[3] Gass, S. F., "Evaluation of Complex Models," <u>Computers and Operations Research</u>, Vol. 4, p. 27-35, 1977.

[4] Wood, D. O., "Model Assessment in the Policy Research Process," lecture at the Stanford University Energy Seminar Series, November 9, 1978.

[5] Greenberger, M., "A Way of Thinking About Model Analysis," paper presented at the Department of Energy/National Bureau of Standards Workshop on Validation and Assessment of Energy Models, Gaithersberg, MD, January 10-11, 1979.

[6] Richels, R. G., and D. Kresge, "Third Party Model Assessment," paper presented at the Department of Energy/National Bureau of Standards Workshop on Validation and Assessment of Energy Models, Gaithersburg, MD, January 10-11, 1979.

[7] Baughman, M. L., "Reflections on the Model Assessment Process: A Modeler's Perspective," paper presented at the Department of Energy/ National Bureau of Standards Workshop on Validation and Assessment of Energy Models, Gaithersburg, MD, January 10-11, 1979.

[8] Goldman, N. L. and J. Gruhl, "Assessing the ICF Coal and Electric Utilities Model," paper presented at the Department of Energy/National Bureau of Standards Workshop on Validation and Assessment of Energy Models, Gaithersburg, MD, January 10-11, 1979.

[9] Stauffer, C. H., "Validation and Assessment of ICF's Coal and Electric Utilities Model," paper presented at the Department of Energy/National Bureau of Standards Workshop on Validation and Assessment of Energy Models, Gaithersburg, MD, January 10-11, 1979.

[10] Holloway, M., "The Texas National Energy Modeling Project and Evaluation of EIA's Midrange Energy Forecasting Model," paper presented at the Department of Energy/National Bureau of Standards Workshop on Validation and Assessment of Energy Models, Gaithersburg, MD, January 10-11, 1979.

[11] Lady, G., "Model Assessment and the Policy Research Process: Current Practice and Future Promise," paper presented at the Department of Energy/National Bureau of Standards Workshop on Validation and Assessment of Energy Models, Gaithersburg, MD, January 10-11, 1979.

[12] Sweeney, J. L., "The Energy Modeling Forum," paper presented at the Department of Energy/National Bureau of Standards Workshop on Validation and Assessment of Energy Models, Gaithersburg, MD, January 10-11, 1979.

[13] Cherry, B. H., "Electric Load Forecasting: Probing the Issues with Models," paper presented at the Department of Energy/National Bureau of Standards Workshop on Validation and Assessment of Energy Models, Gaithersburg, MD, January 10-11, 1979.

[14] Stanford Institute for Energy Studies, "Stanford-EPRI Workshop for Considering a Forum for the Analysis of Energy Options Through the Use of Models, Electric Power Research Institute, Special Report EPRI EA-414-SR, May 1977.

[15] Belden, R. D., "Report on the Forum Project: FOSSIL1 Analysis of Future Trends in the U. S. Coal Industry," in Energy Modeling Forum, "Coal in Transition: 1980-2000," Volume 3, Stanford University, September 1978.

[16] Nissen, D., "Impacts of Assessment on the Modeling Process," paper presented at the Department of Energy/National Bureau of Standards Workshop on Validation and Assessment of Energy Models, Gaithersburg, MD, January 10-11, 1979.

[17] Energy Modeling Forum, "Aggregate Elasticity Estimates," Working Paper EMF 4.4, Stanford University, September 1978.

[18] Krohn, G. C., J. Mehring and J. C. VanKuiken, "The Representation of Markets in Optimization Models," Appendix H in Energy Modeling Forum, "Coal in Transition: 1980-2000," Volume 2, Stanford University, September 1978.

A WAY OF THINKING ABOUT MODEL ANALYSIS

Martin Greenberger

The Johns Hopkins University

Introduction

In the course of writing a review article on the assessment of energy policy models, I have been thinking a good deal about ways of classifying and describing the widening array of activities devoted to the study and investigation of models, as opposed to their development and use. With less frustration and arbitrariness than I might have expected in such an exercise, I finally arrived at two organizational schemata as explained in the review article (1). What I would like to do here is present the rationale for these schemata and show how it helps in reflecting on the routes being taken by model analysis and on where it is headed.

Please note that I am using the term "model analysis" and not "model assessment." For me, the first term includes the second, but does not have as specifically evaluative a connotation. The field of energy policy modeling is young and its growth has been lively, to say the least. Table 1 lists a sampling of the models and their uses, taken from an article written three years ago (2). Some of the most important models entering into energy policy today are not present there. With a field as green and lush as this one, there is a need for mowing and weeding. It is necessary to critique, and umpire, and evaluate. But it is also important to compare, understand, and explore. I employ the term "model analysis" to include all of these activities.

My two organizational schemata for model analysis are displayed in the form of a tree and a table. The tree is shown in Figure 1, the table -- a 4x3 matrix -- in Table 2. The tree has two main branches, corresponding to two principal modes for conducting model analysis, each with its own style and distinguishing set of objectives. The table focuses instead on who has taken the initiative for doing the analysis and under what circumstances. The tree and the table are two different perspectives for viewing the same array of activities. To avoid confusion, they are best dealt with separately. We shall discuss them one at a time.

Table 1

Energy Policy Models, Methodologies, and Uses

Model	Supply Side	Demand Side	Uses
Adams-Griffin	Optimization	Econometric	Strategies for oil refinery pricing.
Baughman-Joskow	Optimization	Econometric	Energy-economic effects of nuclear moratorium in California.
Bechtel Supply	Accounting	Exogenous	FEA studies of industry requirements for energy expansion.
Brookhaven	Optimization	Exogenous	ERDA evaluation of alternative energy technologies.
Coal 1	System dynamics	System dynamics	Congressional hearings on energy forecasts.
DRI-Brookhaven	Optimization	Econometric	ERDA studies of economic impact of alternative energy futures.
Dupree-West	Exogenous	Exogenous	Department of Interior long-term energy forecasts.
Emergency Energy Capacity	Optimization	Exogenous	Office of Energy Preparedness and Treasury Department storage option studies.
ETA and ETA-Macro	Optimization	Informal econometric	Studies of nuclear alternatives (Ford-Mitre, Committee on Nuclear and Alternative Energy Systems).
FEA Short-Term Petroleum	Optimization	Econometric	FEA studies of oil embargo.
Hudson-Jorgenson	Econometric	Econometric	Impact of reduced energy consumption on the economy (Ford, EEI).
Hnyilicza	Econometric	Econometric	Alternative strategies for optimal economic growth.
Illinois Input-Output	Econometric	Exogenous	ERDA studies of energy conservation.
Kennedy-Niemeyer	Econometric	Econometric	Macroeconomic effects of a nuclear moratorium in California.
Lawrence-Berkeley	Optimization	Partial optimization	EPRI industry studies.
MacAvoy-Pindyck	Econometric	Econometric	White House analysis of gas deregulation.
Nordhaus Bulldog	Optimization	Econometric	Energy economic impact of alternative nuclear and fossil fuel strategies (Committee on Nuclear and Alternative Energy Systems).
PACE	Optimization	Exogenous	Energy sector studies with emphasis on petro-chemical industry.
PIES	Optimization	Econometric	National energy plan and FEA studies of oil and natural gas price decontrol.
PILOT	Optimization	Partial optimization	Exploration of potential energy-economic growth.
SEAS (House)	Exogenous	Exogenous	Economic and environmental impacts of alternative energy futures.
SRI-Gulf	Process representation	Informal econometric	Gulf Oil Co. and White House decisions on synthetic fuels.
TERA	Optimization	Econometric	American Gas Association natural gas studies.
Wharton	Econometric	Econometric	Congressional hearings on Carter energy plan.

Glossary

Accounting: Charts of requirements and characteristics displaying numerical relationships.

Econometric: Mathematical (difference) equations solved simultaneously, with coefficients estimated statistically from historical data.

Exogenous: Given or assumed, rather than calculated (endogenously) within the model.

Optimization: Determination of "best" solutions by means of algebraic procedures.

Process representation: Description of energy processes and markets in the form of a hierarchical network.

System dynamics: Mathematical (integral) equations solved recursively with coefficients estimated judgmentally from the modeler's experience and intuition.
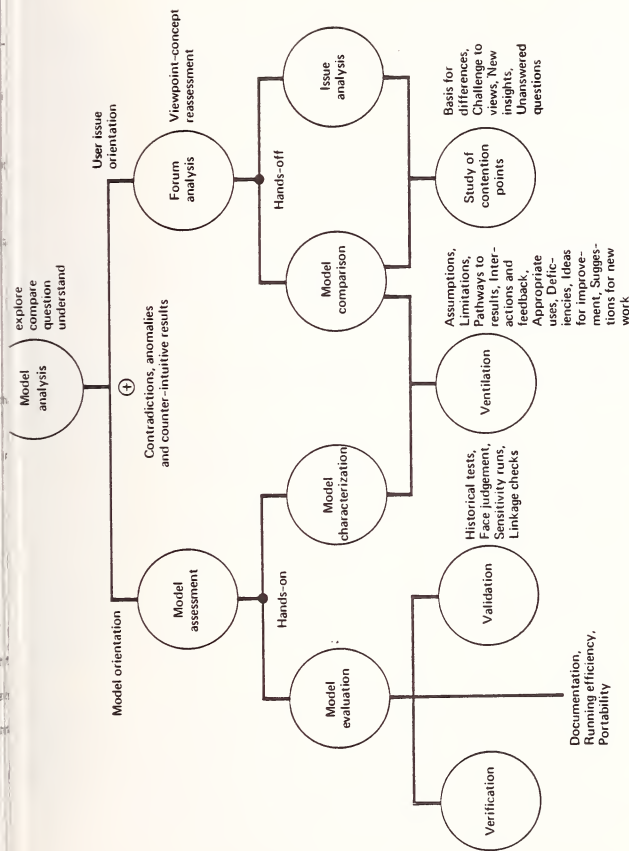
FIGURE 1

Two Kinds of Model Analysis: Their Overlap and Complementarity

301

Table 2
Two-Way Classification of Model Analyses

	NATURAL	AD HOC	INSTITUTIONALIZED
First party / Model developers	Model creation	Modeling groups and workshops	Self–assessment and standards
Second party / Model users	Staff work	Consultants' reports to user	Ongoing user review
Third party / Model analysts	Spontaneous peer review and dissertations	Organized review	Model assessment laboratories
Joint effort / Mixed group	Marketplace of ideas	Discussion meetings and studies	Ongoing forum

The Tree

Let me begin with an admission. I tend to view model analysis from my own backyard. I was fortunate to have the opportunity while on leave from Johns Hopkins University during 1976 and 1977 to put into action some recommendations coming out of a study I had conducted with others the three years before. We had made our recommendations in the setting of an observation on the present sociology of the modeling field.

> "Modelers mostly build and run their own models: that is
> where the credits lie. Very few modelers run and analyze
> the other fellow's model in any systematic way.... Modelers
> are synthesizers and refiners more than analyzers, particularly
> analyzers of other modeler's models. When possible at all,
> such secondary analysis is too difficult and unrewarding an
> activity to generate much interest. As a result, the inner
> workings of a policy model are seldom understood by anyone
> but the builders of the model (and not always by them).
> This is a weak foundation for gaining the reliance and
> trust of policymakers." (3, p.339)

We did not suggest that model builders be made to pay more attention to the models of others. We felt it was enough for them to do their own jobs well. What we did propose was

> "development of a new breed of researcher/pragmatist -- the
> model analyzer -- a highly professional and astute practioner
> of the art and science of third-party model analysis. Such
> analysis would be directed toward making sensitivity studies,
> identifying critical points, probing questionable assumptions,
> tracing policy conclusions, comprehending the effects of
> simulated policy changes, and simplifying complex models
> without distorting their key behavioral characteristics." (3, p. 339)

I took my leave at the Electric Power Research Institute where I served as manager of the Systems Program with a significant budget for sponsoring outside research and a charge to initiate a program of research germane and useful to the Institute and its electric utility membership.

In early 1976, it was already clear that policy models were attracting the attention of energy decisionmakers and regulators in Washington and state governments. There was a partly progressive, partly defensive readiness to accept the argument that projects designed to understand and objectively assess the use of these policy models would be in the interests of the electric utility industry. With the help of my staff and several key colleagues outside of the Institute, I was able to arrange for the establishment of a number of efforts directed at the analysis of energy policy models. Indirectly, these efforts were intended to promote an awareness of the need for model analysis and to stimulate the interest and development of people who would be qualified and able to carry on this kind of activity.

Two of the efforts set up at the time were organized on a trial basis as prototypes and seedlings for possible continuing activities. One was the Model Assessment Group set up at the Massachusetts Institute of Technology, the other was the Energy Modeling Forum established at Stanford University. Both have prospered and both are in high gear at the present time. They are my paradigms for the two branches of the tree in Figure 1, which I refer to as Model Assessment and Forum Analysis.

Having made this admission of unabashed subjectivity in the design of my first classification scheme, let me express the belief that a wide variety of other very important model analysis endeavors fit comfortably in these same two categories and are in no sense diminished by the association. Of course, the fit is never perfect. Some activites will find they have significant elements in both categories. A case in point, again drawn from my own experience, is the Utility Modeling Forum which we were just beginning to put together at the Electric Power Research Institute when it was time for me to return to Hopkins. But more about it later after the two classifications have been described.

The overall process of model analysis is one of probing, exploring, comparing, questioning, and understanding, as indicated at the top of Figure 1. In these ways, model assessment and forum analysis are the same. Their main point of difference comes in the object of the questioning and understanding. Model assessment, in taking the measure of the model, focuses primarily on the model as its object of analysis. Forum analysis focuses more on the issue to which the model is being applied.

In both forms of model analysis, there are contradictions and anomalies occurring in the results that are counter to the intuition of the analyst. Sometimes the problem lies with the model and leads to a correction or improvement. Other times the problem is in the head of the analyst -- with the conceptual model or mental image the analyst carries cognitively of the system being modeled. Then, assuming open-mindedness, it leads to a reassessment of idea or belief. I am deliberately adapting the words "model" and "assessment" in portraying the process of forum analysis to exhibit a parallelism with model assessment. In one case the model is highly structured, explicit, and runs on a computer. In the other case the model is loosely formed, implicit, and resides in the mind of the analyst, user, or modeler. Both models are imperfect. Both need to be examined and adjusted. Either one may be modified as a result of a contradiction taking place at the circled cross in Figure 1. The relative number of times one is modified rather than the other is a measure of the (policy) model. It is a measure of its state of development, and it is a measure of its effectiveness as a learning tool.

Another parallelism exists between model assessment and forum analysis. In a standard model assessment, one model at a time is subjected to scrutiny. A series of computer runs is made to explore and assess the model's capability, limitations, adequacy, and realism (4). One model is run over several issues, in a manner of speaking. In forum analysis, on the other hand, a number of models are applied to one set of questions relating to a single subject of inquiry (5). That is, many models are run on a single issue.

Model assessment examines a model over the full range of its performance. Forum analysis, in contrast, concentrates on a focused set of questions, applies a set of models to these questions, compares the results, and probes the differences. Model assessment is designed specifically to evaluate a model and understand it. Users are not a necessary part of the operation, although they may prove helpful. In forum analysis, they are essential. The aim of forum analysis is to broaden user insights and understanding of the issues as well as the models. As a by-product, modelers who participate along with the users in the forum process gain a fuller awareness of what their models can and cannot do. Forum analysis provides a comparative commentary on the models it employs. But evaluation of these models is not its main purpose.

Figure 1 indicates that model assessment is conducted in the "hands-on" mode while forum analysis proceeds with "hands-off." This is, in fact, only partly true in that certain less intensive forms of model assessment also are performed with hands-off. What this means is either no model runs, as such, are made, or if they are made they are made by the modelers and not the analysts (ordinarily under the guidance or supervision of the analysts). In hands-on operation, the analysts make the runs themselves.

Model assessment and forum analysis are complementary. Their differences are a matter of degree and style. Each has its own set of objectives and each contributes in its own way to increased understanding of models and issues -- model assessment more to the former, forum analysis more to the latter.

Model assessment, as represented in Figure 1, has two aspects: the first is evaluative in nature; the second, called "Model Characterization," is primarily nonevaluative in nature. Evaluation includes verification (assuring that the modelers have followed through faithfully in executing their design plans), validation (checking that the model captures the essence of the real system it attempts to depict and that the data used in development of the model are appropriate, adequate, and accurate), and quality control of the usability of the model and its readiness for use (clarity and comprehensiveness of documentation, cost and convenience of running the model, efficiency of execution, portability, difficulty of obtaining or projecting data, and so on). Improvement of model documentation, almost always a problem, can be an important secondary benefit of a model assessment, as the experience at M.I.T. is showing.

The nonevaluative aspect of model assessment has to do with understanding and characterizing the model and its properties. It includes investigations into the assumptions and limitations of the model, its appropriate uses, and why it produces the results it does. A full comprehension of the model's properties leads naturally to discovery of ways the model can be revised, corrected, extended, simplified, decomposed, linked with other models, and generally improved and made more useful. The ideas and insights coming out of this phase of the inquiry can be one of the most productive outcomes of a model assessment.

For a discussion of the different types of validation used in model assessment, and their relative advantages and drawbacks, the reader is referred to the previously mentioned review article (1). Also given there is an account of the first major assessment undertaken by the group at M.I.T., some of the problems faced in that effort, and some of the lessons learned.

Forum analysis, as portrayed in Figure 1, also has two functional branches. The first is concerned with the comparative runs made of the set of models applied to the issues under investigation. A purpose that model comparison shares with the characterization function under the assessment process is to determine by poking the model why it produces the results it does. Having other models with which to compare, and specific issues on which to focus, gives content to the quest, but by no means makes it simple. "Ventilation" is a term coined to signify this airing of the model (6). Different in purpose from both verification and validation (the two other v's), ventilation opens the model to "sunshine" and close scrutiny, examining its assumptions, limitations, deficiencies, feedback effects, and surprising results, all with a critical eye.

A second goal of model comparison is to help clarify the policy issues on which the analysis is targeted. This goal model comparison shares with the other aspect of forum analysis, called issue analysis. Of special interest are disagreements, known as contention points, whose basis can be explored with the assistance of the model runs. In the best case, it is here that views get challenged and insights deepened. Questions that cannot be satisfactorily resolved are identified as possible topics for further study.

In summary, model assessment and forum analysis both focus on specific selected issues. It makes little sense to analyze a model in the abstract, divorced from the concrete uses for which it is intended or can be applied. Model assessment sweeps over many issues and possible applications in order to understand and evaluate the model. Forum analysis, employing several models, concentrates on a single set of issues in order to frame and illuminate these issues. The differences between these two kinds of model analysis have to do with style, emphasis, and primary objectives. They are complementary activities and share many of the same techniques of analysis.

Table 2 presents the second organizational schema for model analysis in the form of a 4x3 matrix. Each row of the matrix represents a main party to the analysis, either the instigator or the agent. Each column describes the motivation and setting for the analysis.

In the first instance, modelers should be analyzing their models themselves. A model is a creative synthesis of its developer's insights and conceptualizations. Just as authors edit and revise their written drafts during the writing process, so model developers must test and refine the expression of their ideas during the modeling process. Modelers are the "first party" in model analysis. They are depicted by the first row of Table 2. The analysis they perform of their own models serves to sharpen their insights and correct their misconstructions. Own-model analysis, unfortunately, is often abbreviated because of the modeler's understandable impatience to achieve a working version and the feelings of elation and accomplishment that come with a model's finally producing plausible results.

The "second party" in model analysis are the users of the model. They are represented by row 2 of Table 2. Users have a natural incentive to want models examined closely as a basis for choosing one to fit their application, and also in designing runs for it, monitoring its operation, and reviewing its results. Users may employ or call upon modelers to help in the review and selection process. But because of their special relation to modelers whose work they have funded, and because of their possible preoccupation with the modeling results, users are not often in a favorable position for doing a truly objective job of model analysis.

It is with the emergence of the "third-party" analyst that the analysis of models comes into its own. Represented by row 3 of Table 2, third-party analysts are relatively new to the scene, but they do now exist. A growing number of highly skilled practitioners have been developing during the past few years, some already to a point of extraordinary competence. Organizationally, by allegiance, and by incentive, they are detached from both model users and developers. They are set up to perform a role in the policy process that is much needed. It has two sides, described prospectively a while back in the following way:

> to "make policy models more familiar, afford them greater longevity, and give policymakers a place to turn for impartial analysis and assessment." Also, to "produce de facto standards of performance for model builders," to "stimulate (and require) improved, open documentation of models and data," and to "promote a generally higher level of professionalism in the modeling trade" (3, pp. 339-40).

The relationships that third-party model analysts establish with model users and developers are extremely important. Sometimes the three parties will come together to work jointly (if not always harmoniously) reviewing a model as a group. These joint efforts are represented by the fourth and final row of Table 2.

The three columns of Table 2 distinguish among model analyses by the nature of the analysis and the setting in which it arises and is conducted. An analysis can be either "natural" or spontaneous (the first column), occurring as a normal part of one's work assignment without the need for special arrangements; "ad hoc" (the second column), as when a contract is let or a group organized expressly to perform the analysis; or "institutionalized" (the third column), when the analytic activity is established on a long-term, continuing basis.

The schema given in Table 2 provides a structure and logic that are convenient for classifying model analysis activity. Illustrations for each of the cells come readily to mind. The examples presented in the review article derive from personal experience and knowledge, but are easily replaced with many other possible examples from the reader's own orbit of familiarity.

Moving from left to right along each of the four rows in succession, here are the kinds of examples that seem to me to fall into the respective cells of the matrix. Fuller specification is given in the review article.

1,1 Model Creation. Model analysis done as a normal part of the process of constructing a model and in the course of verifying it and checking its validity.

1,2 Modeling Groups and Workshops. The getting together of modelers in groups to discuss policy questions of joint interest and to use their models to address these questions. Common sets of assumptions and scenarios are agreed upon by the group, and results are compared.

1,3 Own-model Assessment and Standards. The possibility of a requirement that developers of models for use by the government complete a questionnaire on the quality and performance of their models that would necessitate much more extensive testing of models by their developers than is customary at the present time.

2,1 Staff Work. Work performed by a user's staff in selecting a model, finding suitable data for it, designing model runs, monitoring results, and suggesting improvements.

2,2 Consultants' Reports to Users. Assistance provided to a user in reviewing models, often in the form of surveys or assessments prepared by a consulting firm or outside research organization.

2,3 Ongoing User Review. A mechanism established by a user or community of users to provide it with a regular review of models and modeling studies of particular interest.

3,1 Spontaneous Peer Review and Dissertations. Independent critiques by persons who are neither direct users of a model themselves nor commissioned by users to analyze the model. They perform their analysis on their own initiative -- for example, as a normal part of the process of peer review, or in the course of writing a graduate student dissertation.

3,2 Organized Review. Prearranged independent assessments organized to supplement or focus spontaneous peer reviews and student dissertations.

3,3 Model Assessment Laboratories. Independent model assessments established as recurring activities and the continuing work of a permanently funded facility.

4,1 A group of model developers, users, and analysts coming together to bring their respective talents, insights, and perspectives to bear in an interdisciplinary multi-faceted endeavor or interchange, often informally and without any concerted measures being taken to rivet attention or develop a consensus. The open marketplace of ideas is the most general form.

4,2 Discussion Meetings and Studies. An organized and focused microcosm of the marketplace of ideas in the form of an assessment study involving model developers, users, and analysts.

4,3 Ongoing Forum. The institutionalization of joint efforts by modelers, users, and analysts to provide a forum kind of model analysis on a continuing basis within the framework of a long-term funded activity with an advisory structure and permanent staff.

So much for Table 2. I have found it a very useful and workable means for classifying model analysis activities and I had no trouble finding examples for each of the cells. But it is not without its gray areas and ambiguities. As a case in point, consider the Utility Modeling Forum (UMF) set up by EPRI's System Program in 1978 to create an ongoing process of comparative modeling analysis and structured discussion of utility problems by members of the utility industry. The UMF intends to concern itself specifically with models and issues of immediate relevance to electric utility companies, and seemed to me to be a perfect choice for inclusion in cell 2,3. But the UMF includes modelers from the utility companies and an argument has been made that it therefore belongs in cell 4,3 along with the Energy Modeling Forum, after which it was originally patterned. Clearly, the definition of cells requires a more legally astute mind than mine to avoid or resolve such questions of class membership.

Conclusion

My observation of the development and performance of model analysis activities within the past few years has reinforced my belief that these activities have a central role to play in making policy models more useful and understandable. It seems very likely to me that model analysis in some form will become a permanent and important component of policy studies. But the shape it will eventually take is not yet clear.

I have recently heard a number of different views expressed on the probable future of model analysis, ranging from the very cynical to the very optimistic. But there was one characteristic they all had in common -- a general consensus that model analysis in whatever form was envisioned was needed.

Some argue that modelers are in the best position to do assessments and that in the long run the development of model analysis will not take the form of a separate discipline; e.g., the performance of assessments by modelers themselves on their own work and the work of others. This is a possibility, although it raises questions of impartiality and it is not clear that modelers will ever wish to spend the very significant amount of time inspecting the work of others that assessment requires. It is also not obvious what would then happen to the forum kind of analysis. But who knows? The practice of policy modeling could change drastically, and if this were the direction it took, so much the better.

Others believe that model assessment and forum analysis will eventually grow to resemble each other more than they do today and ultimately merge into a single type of analysis. This is another possibility, and the current work of the Energy Modeling Forum is indeed moving somewhat in this direction. But each of the two kinds of model analysis offers its own set of advantages and, at the present time, the division of labor is a productive one.

Still another possibility is that the users of models will gradually take more initiative themselves in performing careful assessments of the models they use or are considering. They could join with modelers to do forum analyses on their own, as the Utility Modeling Forum is doing. But here again, we face the question of whether full objectivity could always be preserved. Also, many users may not choose to augment their forces with the numbers of highly skilled, technical people it would take to do the analyses well. They may prefer to continue to look outside for these skills and for performance of the analytic function.

My guess is that the future development of model analysis
will follow several of these alternate paths simultaneously --
including the main path of separate growth. I think that model
analysis will gradually broaden in scope to include not only
comparative (as well as single) model assessments, but policy
study critiques also, where the post mortem is directed not at
a model, but a total policy study within which a modeling effort
may or may not be contained.

Just about everyone has something to gain from model
analysis. The modeler learns more about his model and receives
ideas for improving it and its documentation. The user obtains
a better understanding of models and a firmer basis for making
model selections and interpretations. The model analyst gets
exposure to a broad range of modeling issues and deepens his
perceptions about the modeling process. All three groups
expand their knowledge of the systems being modeled and the policy
issues to which the models are applied.

But the gain is not without cost or pain. The modeler
must endure the pain of criticism and questioning. The user
must suffer the pain of listening to lengthy technical discus-
sions and being subjected to modelers' jargon. The analyst
must accept the uncertainties and insecurities of working in
a new field still without its own reward structure and pro-
fessional recognition.

I asked myself how the gains and pains of model analysis
might net out. To answer the question, I wrote down the two
"gain-pain boxes" shown below, one for model assessment, the
other for forum analysis:

(M.A)	gain	pain	net		(F.A)	gain	pain	net
modeler	+2	-1	+1		modeler	+2	0	+2
user	+1	0	+1		user	+2	-1	+1
analyst	+1	-1	0		analyst	+1	-1	0
total	+4	-2	+2		total	+5	-2	+3

The numbers I have inserted in the boxes are very rough and
subjective, and not very reliable. I might want to change
them tomorrow. Still, they suggest to me that the analyst
has the least net gain at the present time. The future of
model analysis may depend largely on whether the field can
be made more attractive and secure for analysts as it develops
further.

One possible way of making the field of model analysis
more interesting for analysts might be to build rotation into
it. Analysts would serve also as model developers and model
users at earlier stages of their training, before becoming
model analysts, and then perhaps recycle back through this
metamorphosis one or more times during the course of their
careers. This would not only add variety to the profession,
but would strengthen the qualifications of model analysts
and make them much less subject to the criticism that they
do not really understand what modeling, the political issues, or the
problems of the users are all about.

An interesting question is whether model analysts will be well enough treated and their work well enough received by policymakers and model users generally to lead to an institutionalization of their function. This could take the form of a recognized professional discipline with laboratories and centers in many application areas in addition to energy policy, and in many regions of the country. I believe this, too, is a possibility.

For its part, the Electric Power Research Institute is encouraged by its early experiences in promoting the model analysis development. Its Board approved a hefty increase in the level of funding of the model assessment activity in 1978, and its Systems program subsequently circulated a major Request for Proposal for the establishment of an ongoing Model Assessment Laboratory. The Department of Energy, other federal agencies, several large private firms, and a number of additional potential sponsors have expressed interest in financing similar activities, with the Institute or on their own. It all adds up to some interesting times ahead for model analysis, with government policymaking and the public at large as the ultimate beneficiaries.

References

1. Martin Greenberger and Richard Richels, "Assessing Energy Policy Models: Current State and Future Directions," Annual Review of Energy, Volume 4, 1979, forthcoming.

2. Martin Greenberger, "Closing the Circuit between Modelers and Decision Makers," EPRI Journal, Electric Power Research Institute, Palo Alto, October 1977, Number Eight, pp. 6-13.

3. Martin Greenberger, Matthew A. Crenson, and Brian L. Crissey, Models in the Policy Process: Public Decision Making in the Computer Era, Russell Sage Foundation, 1976.

4. "Independent Assessment of Energy Policy Models: Two Case Studies," Energy Laboratory, Massachusetts Institute of Technology, Report No. 78-011, Cambridge, 1978.

5. James W. Sweeney and John P. Weyant, "The Energy Modeling Forum: Past, Present, and Future," Special Issue on Resource Policy Analysis, Journal of Business Administration, 1979, forthcoming.

6. William W. Hogan, "Energy Modeling: Building Understanding for Better Use," Second Lawrence Symposium on Systems and Decision Sciences, October 3, 1978.

Appropriate Assessment[*]

S. C. Parikh
Energy Division
Oak Ridge National Laboratory
Oak Ridge, Tennessee

Introductory Remarks

I had a talk planned that included a number of things dealing with
what steps I intend to take in terms of documentation and usage of a
model that I have been developing on the PILOT modeling project at
Stanford. However, since we are behind schedule and since many of
the things dealing with this model, called Welfare Equilibrium Model
(WEM), can be found in a document that I am in the process of
preparing[1], I will make my talk somewhat briefer and concentrate
on the issues related to model assessment. I will do this, however,
not from a perspective of a professional model assessor, or a
phantom politician decision-maker who just lost an election because
he blindly voted in accordance with the recommendation from a
computer run of a model developed by his political foe, or a phantom
politician decision-maker who just lost an election because he
blindly voted following the advice contained in a shoddy model
assessor's report on an output from a reasonable and valid model,
but from my current perspective of a model builder who is concerned
with improving the models and their contribution in the public
policy arena.

Some of the things that I wanted to say on assessment have already
been said before in this workshop, now about one and a half days
old. But, at the same time, some of the things that I would not have
said have also been said, and therefore, it would appear that it is
somewhat useful to add my somewhat sketchy remarks.

My talk is divided into two parts. First, I would like to make four
introductory remarks that are very much on my mind today, and I
would like to share them with you. Next, I would like to present the
key point of my talk, which is, a concept of appropriate assessment.

My first remark, consisting of assorted but related observations,
has to do with the contribution and role of model assessment.

Listening to the professionals in the field of assessment, the impression I get is that there is a lot of product development of taxonomy. It would appear that adding to and expanding the existing taxonomy, including rhetorical overtones for some of the choices, is the way to understand how one can improve the understanding of the models.

On this score, I heard a talk by Bill Hogan[2] a couple of months ago in which he made a statement that "analysis of analysis is a growth industry". Yesterday, I heard Dave Wood say something to the effect that this industry has experienced a rapid growth, and might experience an equally rapid decline. He also talked of internalization of assessment. If you consider the matrix that Greenberger put up on the chalkboard a little earlier today, one might think of internalization of assessment as reducing research activity in the cell (3,3) (consisting of third party, institutionalized assessments) and increasing the activity in cells (1,1) (consisting of first party or modeler initiated assessment in an uninstitutionalized framework) and (1,3) (consisting of first party or modeler initiated assessment in an institutionalized framework).

All of these assorted observations are leading to the point of my remark, "In relative terms, should the trend be less towards model development and use, and more towards model assessments and assessments of model assessments?" If we have a workshop four years from now, will that workshop focus on modeling and its contribution to understanding of issues, or will that workshop be on assessments of the assessments that were done a few years ago?

The second point I would like to make arose, I am quite sure not for the first time, during the informal discussions at coffee break yesterday. Alan Goldman, Roger Glassey, I and couple of others were having a coffee chat, and one of us, I believe it was Goldman (I stand corrected if he didn't, and take the blame myself), who commented that, in any organization, the complexity of a model just goes beyond the point of manageability. I would like to add two more observations: first, that very few organizations are capable of building complex models, and second, that in depth assessments, the MIT Assessment Laboratory type, because they are costly, can be performed only on few models. Does this mean that we are headed towards fully assessed, complex, large-scale models that are unmanageable and therefore cannot be extensively used, even though they are credible?

Third of my introductory remarks has to do with a working definition of a large-scale model. Again I go back to that coffee conversation that included an idea. We talked of a large-scale model as the one that is large enough to allow one person to develop it, operate it, and use it, either independently or for one or more users. Some help

experts from varied disciplines during model formulation and
ring initial stages is okay, but this concept of "pushing the
mits of one person in managing it" is perhaps a very useful
ncept to think in terms of a large-scale model that is usable.

ing this working definition of a large-scale model, one might
ink of a usable large-scale modeling system as a collection of
delers and models, each modeler operating a model, and the system
nctioning in response to a particular inquiry by an appropriate
bset of modelers collaborating to produce quantitative analyses
eratively by each modeler producing outputs using a given set of
puts, modelers exchanging tables of numbers that revise inputs for
e next iteration, generation of the next round of outputs on the
sis of revised inputs, until a satisfactory intermodel
rrespondence is achieved.

fourth introductory remark has to do with modeling as a way
r quantitative analysts or technicians to effectively participate
the political process. Political process has, by and large, been
accessible to the technical groups, and modeling provides a
hicle for this involvement.

th these introductory remarks, let me move on to the key point that
would like to make in the remainder of my talk.

ing Analysis in Public Decision Making

Exhibit 1, I have attempted to draw a schematic to conceptualize
at I have in mind when we say 'using quantitative analysis in
cision making'. At the center, we have decision makers, planners,
gislators, etc. They receive inputs from many different sources,
ch as their constituents, lobbyists, etc. Quantitative analysis
rms one of such inputs. These inputs mold their thinking with
gard to the problem at hand in order to aid them in developing
ans, reaching decisions, or deciding on their vote. More often
an not they have staff assistants who have the responsibility to
alyze and evaluate these inputs, to identify implications of a
rticular decision, and to develop recommendations on optimal
cision.

the bottom of the exhibit, I have shown a professional group and
s wares. This group, you might say is a group of quantitative
alysts, econometricians, engineers, operations researchers, and
del assessors. The professionals in this group work with some
nformation base. By information base, I mean, raw observations
rom reality as well as transformed data. The transformed data might
e obtained through use of models. The modelers use some of these
ata and produce transformations (in the form of computer printouts
rom models) which are also included in the information base.
cenarios and tabulations produced by the Energy Information
dministration might be viewed as being a part of the information base.

EXHIBIT 1. USING ANALYSIS IN DECISION MAKING
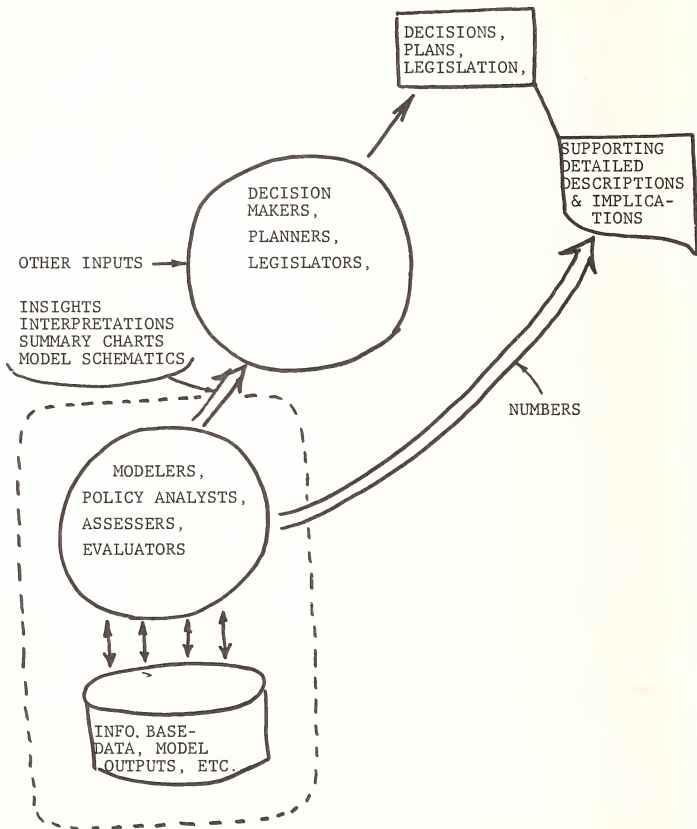
A primary goal of this group of professionals is to provide input into the policy process. Such an input could occcur in one of two generic ways.

The first form of input is through interaction with the decision-makers, the planners, etc. It occurs from analysis of what is in the information base, interpretations, insight development, and their transmittal to the decision-makers through verbal discussions, executive summaries, summary charts, and so on. This is a mechanism through which the decision-makers are better informed, and decisions are made with better understanding and more informed judgement.

Once decisions are made, they are documented and publicized through various reports. The second form of input from the profession occurs in providing detailed numbers that get included in these reports.

There are several things I would like to observe here. First of all, the models and their outputs included in the information base do not themselves produce decisions. The decision-makers do not directly work with information base, but rather, the modelers and the staff personnel of the decision-makers do. On the other hand, however, implications of many decisions may indeed be presented through lots of numbers and tables, thereby sometimes creating an impression that the models produced the decisions. Second, very often today's planning analysts become tomorrow's planners. This is also a way through which some of the quantitative analysis enters into the decision process. In either case, the model formulations and outputs do not produce decisions but the information passes through a human brain, either from the staff to the decision-maker or a staff member becoming a decision maker in the future, and decisions are reached by a human mind that presumably balances many other factors that are left out due to the necessity to simplify and approximate reality in order to manage and perform a quantitative analysis of key tradeoffs.

The name of the game for the profession of model assessors, model developers, etc. is to improve the quality of and confidence in the flow of the insights and in the flow of the detailed sets of numbers, i.e. in the double-lined arrows in Exhibit 1.

Developing Insights Using Models

In terms of developing insights using models, there are two generic ways that this can be accomplished: model development and model exercises with respect to a specific application. Let me put up something here from material on one of my courses. (The talk included a transparency of an illustrative PERT chart from a corporate application).

An example of developing insights through model development is the PERT chart. Such PERT charts have often been used. I use it here simply to convey the point that many insights can be developed simply through the act of putting the model together. One need not have any model runs or any outputs to achieve something tangible as far as the decision-makers are concerned.

There are several observations that I would like to make with regard to such a chart. Most of the insights in such applications are probably developed during model development— the sheer act of putting the model together. The parameter estimates are hardly ever obtained in this context through formal econometrics or use of historical data. They are usually judgemental. Often, high, low, and most likely estimates, and probability distributions are used. Monte Carlo simulation is sometimes applied to develop standard errors. In such an application, the model users are probably the best assessors. Also, such a chart is a policy model in the sense that it aids in making policy decisions.

The second area, that of developing insights through model exercises, is most effective when there is a user interaction during scenario development. If the model is used repeatedly, then user feedback, followed by model improvement, followed by the next set of exercises, etc. occurs. This sort of interactive process is, in my opinion, the most effective method of improving the quality of and confidence in the model outputs.

If no user can be explicitly identified, then there is a definite problem. One man's insight could be another man's misuse of the model. Political inclinations also play a key role.

Some of the useful processes that promote propagation of the insights are: deliberations by a working group of modelers and model users on a set of computer outputs from models, model assessment exercises intended to determine why a model is producing the results that it is, etc.

There is social decision problem, if you will. There is only a limited amount of resources (dollars, people, etc.) available to perform quantitative analysis in the decision-making process. Whether we think in terms of making legislation, or strategic planning, etc., a finite and limited amount of resources are available to perform quantitative analysis. Even though we may not know the exact amount that is available, that number exists.

How should one allocate the total budget for quantitative analysis across various data collection, data validation, model development, model validation and assessment, and model application activities? There are three broad categories of activities. First, better data can improve the quality of and confidence in the flow on the

double-lined arrows in the Exhibit 1. Second, repeated use of models also improves the results coming out of them. For such model uses, the kind of documentation that is needed is sufficient to permit the potential user to determine what the model can do, and to permit the model developer to use it after a period of shelving the model. This is a minimum model documentation that is enough for the user to understand what the model is all about and that permits reuse by the modeler. Finally, we have the third category, detailed model documentation and model assessment. The question is, "How to allocate the budget across these activities so that in some meaningful sense the quality of and confidence in the information flow is improved to the maximum extent?".

If you want to think in terms of the matrix that Greenberger produced on the chalkboard, one might ask the same question with respect to that 3 by 4 matrix, "How do you spread the bucks across that 3 by 4 matrix so that there is a maximum benefit derived from the quantitative analysis?"

Appropriate Assessment

There are extremes that one could follow. For example, in a lecture that I attended some time ago on model assessment, the following theme was presented. This presentation was concerned with a model that was not yet built, but that was something that was going to be built. There was a recent proposal to build the model. Three phases dealing with model development and assessment were outlined.

The first phase consists of Venting of the Research Plan. This is where the proposal is looked over by the peers and so on. It is vented and critiqued. Depending upon the feedback, the model formulation may be modified.

The next phase consists of doing research, model development, intense assessment process, detailed documentation during model development, and more assessment.

In the third phase, after the model development is completed (is it ever?), there is selected replication and counter-modeling (i.e. more assessment).

A question was asked by someone in the audience whether all of these phases should be completed before the model is ever used for the first time in the policy process. The detailed answer given essentially amounted to an unqualified yes. "It is only after the model is fully developed and assessed that it should be used in the policy setting," was the answer.

To my knowledge, there does not exist a computer model that has been developed or is under development in accordance with an orthodox

application of the process described above. Therefore, one could only guess at its probability of success. I would venture to guess, and if you recall my early remarks about complexity of the models and exceeding the capability of the organization to manage them, etc., and also on the model assessment being a very expensive process, this approach of complete development, detailed documentation, model assessment and counter modeling before first application, stands an excellent chance of leading to well-documented and intensely assessed but outdated and never-used models. It also stands an excellent chance of the professionals simply talking to themselves.

At the other extreme lie, of course, the idea of unchecked model development by the modeler in isolation, and resulting model that could very well be an extension of the modeler and may not contain even a slightest element of objectivity or rational analysis.

I would like to introduce a concept of appropriate assessment that, I believe, promises a much greater benefit for the public dollar than any of these extremes. You might think of this approach as using different colored folders, each color representing a particular level of maturity of the model. In the early stages of model development, the model is used in the policy setting but only a small amount of confidence is placed in its results. Put the model and the results in a red tagged folder, if you like.

In the middle stages of model evolution, in which the model is repeatedly used and is constantly being improved on the basis of user feedback and improved understanding of the model structure and the relative importance of the modeled cause-effect relations. The results are also given progressively greater credibility and the model "moves from folder to folder". During this period, the model documentation increases to fit the model maturity. Also, internalized assessments are going on. Just as "wine improves with age", "a model improves through usage, feedback, and added usage". During this period, varying degrees of third party assessment could also be undertaken depending upon such needs.

Repeated usage in different applications is a primary and the most effective way through which the quantitative analysis can contribute in the public policy process. Therefore, it is useful to keep in mind that model development, model usage, and model improvement are the activities where the modeling related research budget dollar is likely to provide the greatest benefit. In this context, model assessment is probably best viewed as an essential "overhead" activity that can, along with many other overhead activities, contribute towards identification and channeling of model building activity in most productive directions.

Finally, as the model passes its prime use (by now, we have several models around that are in this category), model historians take over, and assess it to extract all the useful theory in the model and fully document it for possible use in the future generation of models. This is perhaps one of the most useful roles that many of the assessors can play.

References

[1] Parikh, S.C. (1979), "A Welfare Equilibrium Model (WEM) of Energy Supply, Energy Demand, and Economic Growth", Technical Report, Systems Optimization Laboratory, Department of Operations Research, Stanford University, Stanford, California.

[2] Hogan, W.W. (1978), "Energy Modeling: Building Understanding for Better Use" Proceedings of Second Lawrence Symposium on Systems and Decision Sciences, Berkeley, CA, pp. 1-9.

[3] Wood, D. (1979), Presentation at this Workshop.

[4] Greenberger, M. (1979), Presentation at this Workshop.

A DECISION ANALYST'S VIEW OF MODEL ASSESSMENT

Edward G. Cazalet
Decision Focus, Incorporated
Palo Alto, California

INTRODUCTION

As a decision analyst, I must be very skeptical about the usefulness of model assessment and validation for two reasons: First, any assessment activity should focus on the quality of the decision process; the quality of any models used in the process is only one area of assessment. Second, assessment and validation are extremely difficult tasks to do well because of the necessity for hindsight. Despite this skepticism, I shall try to make a positive contribution to this workshop on model validation and assessment by using the framework of decision analysis to outline an approach to assessment. I will begin by first reviewing the basic concepts of the decision analysis framework.

THE FRAMEWORK OF DECISION ANALYSIS

Decision analysis is a term used to describe a professional practice and methodology for aiding decision making [1-10]. The framework of decision analysis is designed to improve a decision process but is also can be viewed as a framework for assessing the quality of a decision process. In non-technical terms, the framework of decision analysis is outlined in Figure 1.

Good Decisions Versus Good Outcomes

The first step in describing the decision analysis framework is to define a decision. A decision is an irrevocable allocation of resources in the sense that it would require a large amount of additional resources to change the allocation.

The next step is to distinguish between a good decision and a good outcome. A good outcome is one that is favorably regarded by those with the power to make the decision. A good decision, however, cannot be defined as one that produces a good outcome. Because of uncertainty, a good decision may produce either a good or bad outcome.

A good decision must be defined in terms of the process of decision making. Loosely speaking, we would like to increase the likelihood of good outcomes by doing all we can to gather information, create new
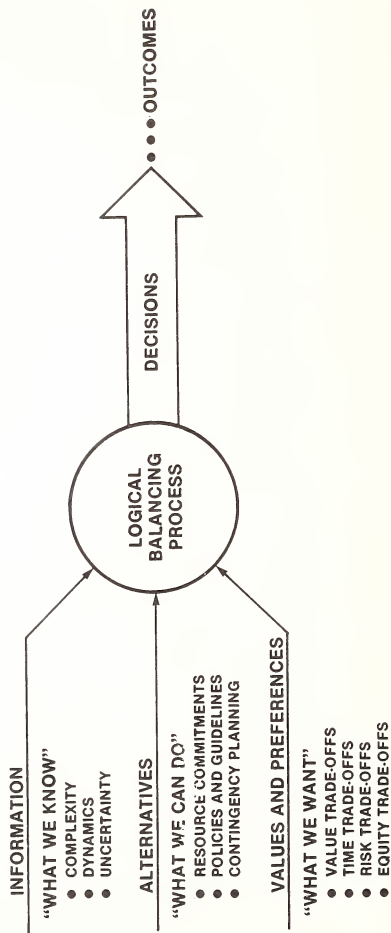
INFORMATION

"WHAT WE KNOW"
- COMPLEXITY
- DYNAMICS
- UNCERTAINTY

ALTERNATIVES

"WHAT WE CAN DO"
- RESOURCE COMMITMENTS
- POLICIES AND GUIDELINES
- CONTINGENCY PLANNING

VALUES AND PREFERENCES

"WHAT WE WANT"
- VALUE TRADE-OFFS
- TIME TRADE-OFFS
- RISK TRADE-OFFS
- EQUITY TRADE-OFFS

LOGICAL BALANCING PROCESS

DECISIONS

• • • OUTCOMES

Figure 1: FRAMEWORK OF DECISION ANALYSIS

alternatives, and contemplate what is a good outcome. But the process of decision making itself consumes resources and time. Therefore, we must define a good decision as one that is the result of a process that economically balances all aspects of the decision problem including the cost of the process itself.

Decomposition of the Decision Problem

The basic idea of decision analysis is to gain insight into complex decision problems using a "divide and conquer" or decomposition approach. We proceed by decomposing a complex decision problem into a number of elements or subproblems, each of which is easier to analyze than the original problem. Then we combine the analyses of the subproblems into an overall analysis of the original problem. Figure 1 shows the decomposition of the decision problem into three basic elements: information, alternatives, and preferences. Each of these elements is analyzed independently and recombined in an iterative process of analysis that is designed to provide a better understanding of the original, complex decision problem.

Information. Information describes "what we know." Information can be represented in two ways: by means of relationships structured in the form of a model and by means of probability assignments.

Structural information consists of information describing how things are connected to other things. For example, we know that the electrical energy produced by an electric power plant is related to its fuel use. Typically, structural information can be represented in terms of equations relating several variables. The value of a quantitative model is greatest when we have many equations and variables. Here the unaided human mind is unable to cope with the solution of many equations in many variables, whereas a computer model can solve thousands of equations [11].

We will later consider the role of structural models in more detail, particularly as their quality relates to the quality of the decision process.

The second way of representing information is by means of probability assignments. There is only one way to communicate uncertainty and that is the language of probability. Decision analysis views probability as a state of mind rather than things. This subjective view of probability includes situations where quantity of experimental data is influential in the probability assignment.

A major area of concern in decision analysis is compensating for the human motivational and cognitive biases that may influence the assignment of probabilities [12-18]. As we shall see, the psychological and analytical techniques that have been developed for assigning probabilities provide a number of useful insights into model validation and assessment.

Alternatives. Alternatives describe "what we can do." It is important that an analysis consider the full range of decision alternatives. Decision analysis is normally thought of as a procedure for selecting among a set of well-defined alternatives. However, an important aspect of decision making is the creative process of generating new alternatives. Often, an analysis will facilitate the creation of new alternatives by focusing attention on the important aspects of the problem. For example, the inclusion of uncertainty in an analysis may suggest hedging alternatives and contingency plans that might otherwise not be considered.

Preferences. Preferences describe "what we want." The importance of preferences from the perspective of model assessment is to identify the important role of preferences in the use of models for analysis. In a decision analysis it is useful to distinguish between four types of preferences: value, time, risk, and equity [19-25].

Value assignment concerns trade-offs between the known consequences of a decision; uncertainty and risk preference are treated elsewhere. In public decision problems, values might be set on the health, mortality, and esthetic consequences as well as the monetary consequences of a decision. Often it is convenient to assign values in monetary terms, but it is not necessary to do so.

Time preference concerns trade-offs between outcomes distributed over time. When values are expressed in monetary terms it is often useful to use a discount rate to characterize time preference.

Risk preference is a term used to describe the fact that most people are not willing to choose among alternatives simply on the basis of the expected value of each alternative (the probability weighted values of all possible outcomes of a decision). Risk preference is therefore a reflection of attitude towards uncertainty; uncertainty itself being described in probabilistic terms in the information element of the analysis.

Equity trade-offs are relevant in decision problems where the outcomes of more than one party are of concern. Equity trade-offs describe how value to one party is to be traded off against value to each other party for purposes of making a decision. In making equity trade-offs it may be difficult or unnecessary to get general agreement among the parties.

Logical Process of Analysis

A decision analysis proceeds by iteratively decomposing a decision problem into its basic elements (information, alternatives, and preferences) and then combining these elements into an overall analysis. At each stage of this process the decision makers or appropriate specialists are involved in developing and analyzing each element. The final result is not so much identification of a good decision as it is development of insight into what makes a good decision. If the analytical process is effective then the intuition of the decision makers should be consistent with the insight from the analysis and the resulting decision

will be logically consistent with the information, alternatives, and preferences of the decision makers.

A good decision of course does not guarantee a good outcome. Most people, however, desire to use logical decision procedures because they believe that these procedures produce the best chance of achieving a good outcome.

When a decision analysis is carried out formally, one of the by-products is documentation of the basis for the decision. The formalism of decision analysis provides documentation of the information, alternatives, and preferences of the decision makers at the time the analysis was carried out. This documentation is of great value in an assessment of the quality of the decision process.

Assessment of the Quality of the Decision Process

Having outlined the framework of decision analysis we are now in a position to consider how we might go about assessing the quality of a decision process.

First, we must emphasize that trying to determine after the fact what was in the minds of decision makers is extremely difficult. While a documented decision analysis will go a long way towards this end, complete documentation of the basis for a decision is not feasible. A good decision rests on many constantly changing subjective elements as the framework of decision analysis makes clear.

Perhaps a more useful way to view assessment is as part of the decision process. The framework of decision analysis provides a structure within which review and testing of alternative modeling assumptions can be accomplished as the analysis is carried out. It is often useful to have skilled "third party" decision analysts and other experts review an analysis at various stages before a final decision is made. The simultaneous assessment of the analysis would reduce documentation needs and increase the chance that assessment could have a positive impact on the analysis. However, only very important problems, or problems that must be solved repetitively, would justify such a high level of attention.

Assessment can best be defined in decision analysis terms by the following two questions: Would additional analytical effort be economic in improving the quality of a decision? Are there areas where too much or too little effort or effort of the wrong kind was applied? In other words, would the decisions that are the focus of the analysis be changed by a different allocation of analytical resources. Note that if we do not identify the decisions that are the focus of the analysis, we cannot apply this test.

Within the framework of decision analysis there are four areas where assessment might be carried out: assessment of the information, alternatives and preferences elements, and assessment of the overall analytical decision process.

In terms of this workshop, it should be noted that model assessment is one aspect of assessing the information element of the decision process. Other people may define a model more broadly covering other elements of the decision analysis framework. I will not quibble with their definition as long as they are willing to distinguish the elements of the decision process within their model or identify the elements not addressed in their model. Assessment is likely to be most useful, however, when it is applied to the entire decision process including any models that were used.

Assessment of Structural Models

Assessment and validation of structural models is what most people mean by model assessment and validation. In terms of verifying that a model is operating as intended and testing the sensitivity of the model results to structural changes there are two major contributions of the decision analysis framework.

First, sensitivity tests and verification of model structure should be measured in terms of the effect on the decisions that the model was designed to address. In this way, the importance of the results of each test can be judged using the most meaningful possible measure.

Second, a final test of a good model is whether the detail and cost of a model are economically balanced. In a good model, it should be hard to single out an area where the model could be greatly improved relative to several other areas.

As an aside, it is important to observe that there is increasing activity in the development of software systems to facilitate and modularize the software development process [26,27]. To the extent that submodels within large-scale models become standardized and easier to understand, the assessment of models will become easier. On the other hand, as we begin to more effectively use the computer in constructing models, it is likely that many more models will be built, each one better tailored to specific decision problems. This proliferation of models will make assessment of models more difficult.

Assessment of Probability Assignments

The basic assessment technique for probability assignments is an integral step in the decision analysis procedure. Using sensitivity analysis, a decision analyst attempts to distinguish those exogenously determined variables in a model that when varied over their approximate range of uncertainty have a major influence on the selection of the best decision alternative. For these variables, their uncertainty needs to be quantified. Other variables simply can be set at nominal values. In an assessment of a decision process one can check to see if sensitivity analysis has been properly performed by redoing it. It is also easy to determine whether crucial variables have been treated as uncertain in the analysis.

A central area of research in decision analysis has concerned the potential biases inherent in the assignment of probabilities by experts. Research by psychologists and others has identified two kinds of bias: cognitive and motivational.

Cognitive biases relate to how people process information. Research has shown that experts tend to think that they know more than they really know. Tests show that untrained experts often assign probabilities of 1 in 100 to events that can be verified as occuring up to fifty percent of the time. As a result there is a danger that an analysis will presume greater certainty than would be presumed if cognitive biases were not present.

Fortunately, research has also shown that training of experts can improve their ability to assign probabilities that authentically represent their true state of mind. Thus one task of assessment is to check whether the expert information used in an analysis has been developed using good probability assignment techniques and proper training of experts.

Motivational biases in probability assignments arise when an expert's beliefs do not reflect his conscious beliefs. A good example is asking an R&D project manager to estimate the probability that his project will be successful when his employment depends on his response. By working with other experts and using proper interview techniques, it is often possible to adjust for the presence of motivational bias. An assessment of an analysis should determine whether the appropriate experts were used in view of the possibility of motivational bias.

Assessment of Alternative Specification

In this area an assessment must consider whether an appropriate effort to include all important alternatives and create new alternatives was carried out. Were some alternatives intentionally left out of the analysis? Did they consider and evaluate information gathering alternatives and hedging strategies?

Assessment of Preference Assignments

Preferences are subjective judgements and therefore particularly difficult to validate. With respect to value, time, and risk preferences for a single individual we can ask whether an appropriate level of effort was expended in structuring an individual's preferences or was some arbitrary characterization of his preferences used. Other tests include checking the consistency in the preference model and its sensitivity to the preference model parameters.

In multi-party problems, we can test whether each party's preferences were considered and whether alternatives such as side payments, where ethical and legal, were considered as a means of generating new alternatives that might be better for all parties.

Assessment of the Overall Decision Process

The basic test here is whether the results of the analysis and the insight of the decision maker coincided at the time the decision was made. Throughout an analysis, insight and analytical results should be continually tested against each other. If the desired end result is not achieved then we should identify the reason. Perhaps the analysts were not sufficiently skilled, or their results were untimely. Or possibly the decision environment changed so that an analysis was no longer necessary or relevant.

Another frequent occurrence is that the decision maker is never really involved in the analysis so that the necessary communication does not take place.

Perhaps the greatest problem in politically important decision problems is that the decision maker is not really interested in an objective analysis of alternatives. He may desire to use an analysis or computer model to advocate a particular position. Or he may simply wish to retain control of the situation because of his fear that decision analysis and modeling will somehow reduce his power.

Conclusions

Decision analysis gives us a framework for performing an analysis and also assessing the quality of a decision process. The primary contribution of this framework is to highlight those aspects of a model or analysis that are relevant to a decision problem so that the importance and quality of all aspects of an analysis can be assessed.

Unfortunately, it is likely to be some time before an assessment process based on decision analysis principles is routinely implemented in the U.S. government, for example. There is no technical reason why these decision analysis techniques cannot be applied on a regular basis to major policy decision problems that justify the efforts required. Their applicability has already been demonstrated in several instances [28-33]. Rather, there is a strong tendency on the part of both modelers and decision makers to avoid making explicit the decisions that are the focus of an analysis. Modelers often prefer anonymity of science whereas politicians prefer to retain full control. In fact, both views are fallacies since the science embodied in decision analysis can address the political aspects of decision problems in a way that enhances the proper role of the political process in gathering information, debating alternatives and resolving equity tradeoffs.

REFERENCES

[1] D. W. Boyd. <u>Decision Analysis: A Primer</u>. Boston: Winthrop Publishers, (forthcoming).

[2] R. V. Brown, A. S. Kahr, and C. R. Peterson. <u>Decision Analysis for the Manager</u>. New York: Holt, Rinehart, and Winston, 1974.

[3] R. A. Howard. "Decision Analysis: Applied Decision Theory." In <u>Proceedings of the Fourth International Conference on Operations Research</u>. New York: Wiley, 1966, pp. 55-71. Reprinted in <u>Readings in Decision Analysis</u>. R. A. Howard <u>et. al</u>. (eds.). Menlo Park, CA: Stanford Research Institute, Decision Analysis Group, 1976, pp. 83-101.

[4] R. A. Howard. "Decision Analysis In Systems Engineering." <u>Systems Concepts: Lectures on Contemporary Approaches to Systems</u>. R. F. Miles (ed.). New York: John Wiley and Sons, 1973. Reprinted in <u>Readings in Decision Analysis</u>. R. A. Howard <u>et. al</u>. (eds.). Menlo Park, CA: Stanford Research Institute, Decision Analysis Group, 1976, pp. 45-81.

[5] R. A. Howard. "Decision Analysis: Perspectives on Inference, Decision, and Experimentation. <u>Proceedings of the IEEE</u>, Vol. 58 No. 5 (May 1970), pp. 632-643. Reprinted in <u>Readings in Decision Analysis</u>. R. A. Howard <u>et. al</u>. (eds.). Menlo Park, CA: Stanford Research Institute, Decision Analysis Group, 1976, pp. 543-556.

[6] R. A. Howard, J. E. Matheson, and D. W. North. <u>Decision Analysis for Environmental Protection Decisions</u>. Final Report prepared by Stanford Research Institute, June 1977. SRI Project 5094.

[7] R. A. Howard. "The Science of Decision-Making." <u>Readings in Decision Analysis</u>. R. A. Howard <u>et. al</u>. (eds.). Menlo Park, CA: Stanford Research Institute, Decision Analysis Group, 1976, pp. 147-164.

[8] J. E. Matheson and R. A. Howard. <u>An Introduction to Decision Analysis</u>. Menlo Park, CA: Stanford Research Institute, Decision Analysis Group, 1968. Reprinted in <u>Readings in Decision Analysis</u>. R. A. Howard <u>et. al</u>. (eds.). Menlo Park, CA: Stanford Research Institute, Decision Analysis Group, 1976, pp. 5-43.

[9] D. W. North. "A Tutorial Introduction to Decision Theory." In *IEEE Transactions on Systems Science and Cybernetics*. SSC-4, No. 3 (September 1968). Reprinted in *Readings in Decision Analysis*. R. A. Howard *et. al.* (eds.). Menlo Park, CA: Stanford Research Institute, Decision Analysis Group, 1976, pp. 103-115.

[10] H. Raiffa. *Decision Analysis: Introductory Lectures on Choices Under Uncertainty*. Reading, MA: Addison-Wesley Publishing Co., 1968.

[11] E. G. Cazalet. *Generalized Equilibrium Modeling: The Methodology of the SRI-Gulf Energy Model*. Final Report prepared by Decision Focus Incorporated for the Federal Energy Administration, May 1977.

[12] C. S. Spetzler and C.A.S. Stael von Holstein. "Probability Encoding in Decision Analysis." *Management Science*, Vol. 22, No. 3 (November 1975). Reprinted in *Readings in Decision Analysis*. R. A. Howard *et. al.* (eds.). Menlo Park, CA: Stanford Research Institute, Decision Analysis Group, 1976, pp. 403-427.

[13] C.A.S. Stael von Holstein. *Assessment and Evaluation of Subjective Probability Distributions*. Stockholm: The Economic Research Institute at the Stockholm School of Economics, 1970.

[14] C.A.S. Stael von Holstein. "A Bibliography on Encoding of Subjective Probability Distributions." Unpublished manuscript, Stanford Research Institute, 1972.

[15] A. Tversky and D. Kahneman. "Availability: A Heuristic for Judging Frequency and Probability." *Cognitive Psychology*, Vol. 5, 1973, pp. 207-232.

[16] A. Tversky and D. Kahneman. "Judgment Under Uncertainty: Heuristics and Biases." *Science*, Vol. 185 (September 27, 1974), pp. 1124-1131.

[17] R. L. Winkler. "The Assessment of Prior Distribution in Bayesian Analysis." *Journal of the American Statistical Association*, Vol. 62 (1967), pp. 776-800.

[18] R. L. Winkler and A. G. Murphy. "Experiments in the Laboratory and the Real World." *Organizational Behavior and Human Performance*, Vol. 10 (1973), pp. 252-270.

[19] K. Arrow. "Aspects of the Theory of Risk Bearing." Yrjö Johnson Lectures, Helsinki, 1965.

[20] D. W. Boyd. *A Methodology for Analyzing Decision Problems Involving Complex Preference Assessments*. Ph.D. Dissertation, Stanford University, 1970. Available from University Microfilms (No. 71-2737), Ann Arbor, Michigan.

[21] D. W. Boyd and C. E. Clark, Jr. Multi-Party Decision Analysis for Social Decisions. Decision Focus Incorporated Working Paper No. 2, August 1978.

[22] R. A. Howard. "Life and Death Decision Analysis." In Proceedings Second Lawrence Symposium on Systems and Decision Sciences, 2:271-277, 1978.

[23] R. A. Howard. "Risk Preference." Readings in Decision Analysis. R. A. Howard et. al. (eds.). Menlo Park, CA: Stanford Research Institute, Decision Analysis Group, 1976, pp. 429-465.

[24] Ralph L. Keeney and H. Raiffa. Decision Analysis with Multiple Conflicting Objectives: Preferences and Value Tradeoffs. New York: John Wiley and Sons, 1976.

[25] von Neumann and Morgenstern. Theory of Games and Economic Behavior, Second Edition. Princeton, New Jersey: Princeton University Press, 1947.

[26] Ronald J. Adler et al. The DFI Energy-Economy Modeling System. Final Report prepared by Decision Focus Incorporated for the U.S. Department of Energy. DFI Project No. 1023, December 1978.

[27] E. G. Cazalet et. al. DFI Model Data Management System. DFI Working Paper.

[28] E. G. Cazalet, C. E. Clark, Jr. and T. W. Keelin. Costs and Benefits of Over/Under Capacity in Electric Power System Planning. Final Report prepared by Decision Focus Incorporated for the Electric Power Research Institute, October 1978. EPRI EA-927.

[29] E. G. Cazalet et al. Decision Analysis of Nuclear Plants in Electrical System Expansion. Final Report prepared by Stanford Research Institute for Comision Federal de Electricidad, Mexico City, D.F., 1968.

[30] R. A. Howard, J. E. Matheson, and D. W. North. "The Decision to Seed Hurricanes." Science, Vol. 176 (June 16, 1972), pp. 1191-1202. Reprinted in Readings in Decision Analysis. R. A. Howard et. al. (eds.). Menlo Park, CA: Stanford Research Institute, Decision Analysis Group, 1976, pp. 291-304.

[31] J. E. Matheson and W. J. Roths. "Decision Analysis of Space Projects: Voyager Mars." National Symposium of the American Astronautical Society, June 11, 1967. Reprinted in Readings in Decision Analysis. R. A. Howard et. al. (eds.). Menlo Park, CA: Stanford Research Institute, Decision Analysis Group, 1976, pp. 309-340.

[32] D. W. North and M. W. Merkhofer. "Analysis of Alternative Emissions
      Control Strategies." Chapter 13 in _Air Quality and Stationary Source
      Emission Control_, A Report by the Commission on Natural Resources,
      National Academy of Sciences, Washington, D.C.: U.S. Government
      Printing Office, March 1975.

[33] Synfuels Interagency Task Force. _Recommendations for a Synthetic
      Fuels Commercialization Program_. U.S. Government Printing Office,
      Stock Number 041-001-00111-3. November 1975

VALIDATION ISSUES--A VIEW FROM THE TRENCHES*

W. Marcuse, F. T. Sparrow,** D. A. Pilati

Economic Analysis Division
Brookhaven National Laboratory
Upton, New York 11973

I.   Introduction

A great deal of attention has been directed towards model evaluation
and assessment. A bibliography compiled by Saul Gass lists 37 articles and
monographs and 14 books and reports devoted to model evaluation or assess-
ment. (Gass, undated) Most of these, in dealing with verification and
validation, discuss means and mechanisms by which "outside" parties can per-
form peer review to provide verification (model behavioral response is as
intended and publicized) and establish the validity (model produces results
one would expect, e.g., in the case of most models, it will recreate history)
of models. (Gass, 1977) Little attention is paid to activities performed
by the user modeling team itself to improve the ability of the model to pro-
vide information useful in the decision making process, and to provide con-
fidence that the information is meaningful.

This paper presents a number of case histories describing our experience
with this type of model improvement activity which we have called internal vali-
dation. Our experiences are illuminating since they were learned in the context
of formulating, developing, and exercising a specific set of process models.
This experience has convinced us that internal validation schemes (our defini-
tion) should be incorporated in the project description and that they be used
in part to answer questions of formulation. Having discovered the need to
perform explicit internal validation, we recommend that modelers incorporate
sufficient funding in their project plans to carry out this function and to
fully document it. In general, this will be an unwelcome addition to sponsors
already unhappy with the size of their modeling budget.
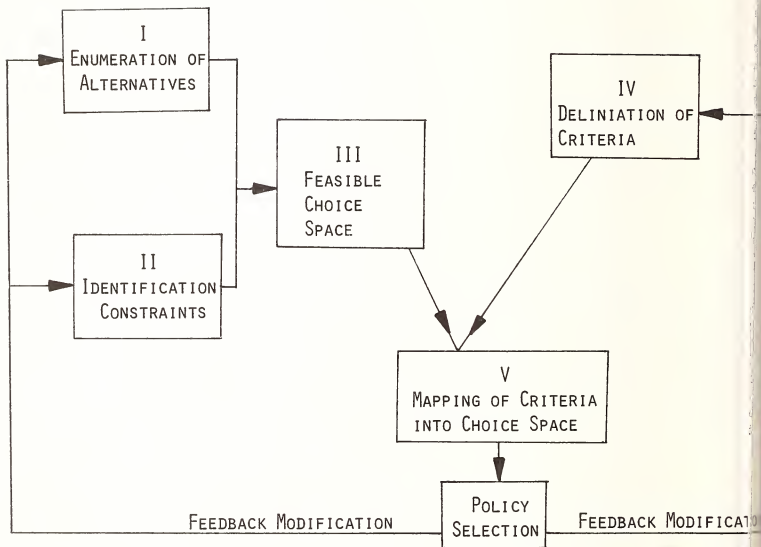
II.   The Decision Process

Increasingly, we turn to government to intercede in areas where economic
equilibrium is subject to market failures, where externalities previously
ignored are now considered socially undesirable, or where political goals
have to be satisfied. These activities require the manipulation of enormous
data bases. This has prompted an increased acceptance of information pro-
vided by quantitative models capable of such manipulation by the actors in
the decision process and an increased demand for such tools. It is not
surprising that model builders and users have evidenced increased concern
with regard to the quality of their products.

Figure 1 indicates where such models can fit into the decision process.
Decision makers are faced with a wide range of policies and actions (Box 1).
They also are acutely aware of the political and institutional limitations

---

FIGURE 1
An Idealized Decision Process

MODELS GOOD AT III, V: DECISION MAKERS MUST PROVIDE I, II, IV; THEN
CLEAR THAT MODELS ARE ADJUNCT TO DECISION PROCESS.

on their freedom to pursue these alternatives (Box II). Physical and economic constraints are also recognized although their perception may be dimmer. The interaction of the alternatives with the identified constraints identifies a "feasible choice space" (Box III). The decision maker must also specify the value system weights that will be used to rank-order the outcomes of alternative policy decisions (Box IV). This is a very difficult and painful task and is often accomplished poorly within the decision structure. Mapping the criteria the choice space yields an ordering of the outcomes (Box V). Conceptually, this process results in the identification of the preferred policy choice. Models are generally recognized as performing the tasks in Boxes III and V quite well; however, the decision makers must provide the bulk of the information required from Boxes I, II, and IV. Models influence these activities only by the feedback loops shown in the figure. Unfortunately, the information provided by the models will never be perfect but hopefully can be improved. It is the process of improvement that we shall call validation. The purpose of models we are examining is to lead to improved decisions and the purpose of model validation is to lead to improved information flow into the decision process. Model improvement occurs not only as the model meets criteria or standards set by a professional modeling community, but as the modeling process better suits the needs of the decision process, i.e., users should play a key role in validation. This does not mean that professionally derived criteria should be ignored but rather that the professional criteria should be developed so that validation is defined within the decision context. Hence, it may differ from topic to topic, model to model, and even from decision maker to decision maker within the same topical area using the same model.

If the object of model validation is to improve the model, then how does one define improvement? One definition might define improvement as model modification which leads to better decisions. Note that this definition requires the term "model" to be interpreted as a complete process including formulation, development, application, documentation, interpretation, and review. In the absence of a meaningful operational measure of "better" decisions an alternative (but still qualitative assessment) of validity might be whether the actors in the decision process feel comfortable with the modeling process and its results.

Two cardinal rules that should be adhered to in policy modeling activities are:

1. Users (decision makers or their staffs) should participate in the entire modeling process including frequent review during the development phase since reformulation is a continuing activity.

2. The choice of key variables and the model structure must be consistent with the key policy questions faced by the decision maker. The assumptions, strengths, and limitations of models and their results must be clearly understood if models are to be used effectively in the decision process. (Greenberger, 1976)

To some extent, asking decision makers and their staffs to participate in modeling activities is unrealistic. However, unless considerable interaction takes place, especially with respect to model formulation and the interpretation of outputs, not only are models likely to be ignored, but worse, the product may be used improperly. Introducing a process which insures that the model behaves as intended and that there is agreement between the behavior of the model and the real world will do nothing to insure that the model is designed to answer key user questions or that the model assumptions and limitations are fully acknowledged by the user in the interpretation of the model outcomes.

Finally, it is important to recognize that models which might stand up quite well in comparing the difference in outcomes under alternative policies might fare quite poorly as simulators of history whereas models that simulated past history well might mask or accentuate the effects of alternative policies. (Marcuse, 1979) We must make sure to avoid this trap when attempting to use the ability of a model to reproduce history as a validation criterion; more will be said about this problem later in the paper.

III. The Modeling Process

The modeling process consists of model development, application, and internal validation and the feedbacks associated with these activities. These subprocesses are inseparably intermeshed. The nature and content of the development and application subprocesses are clear. The character of the validation subprocess is obscure, often unrecognized, and seldom documented. Our experience indicates that it is critically important, and that not only should internal validation be explicitly incorporated as a task in a modeling effort but also that modeling efforts should specifically require documentation of internal validation results.

Figure 2 presents a functional breakdown of policy modeling activities. The modeler occupies the central box on the diagram. He develops, exercises, and improves a model that combines data (III) and causal relationships (II) to provide answers to key questions posed by the user (I). Because decision models by their nature span several disciplines (e.g., economic, engineering, environmental), the modeling team should be multi-disciplined so that causal relationships from the various disciplines are correctly specified and data are properly interpreted.

The link between the modeler and the user calls for the modeler to ascertain jointly with the user what information the user needs in response to what questions the user might ask. In passing, it might be noted that some have attributed the demise of RANN in NSF to its failure to properly recognize the need for close co-operation between users and modelers. The modeling effort must be closely integrated with those involved with the policy planning process in order to assure pertinent and useful information. (NAS, 1976).

FIGURE 2

THE MODELING PROCESS: MODEL STRUCTURE A COMPROMISE--
BRIDGES GAP BETWEEN DATA AND NEEDS
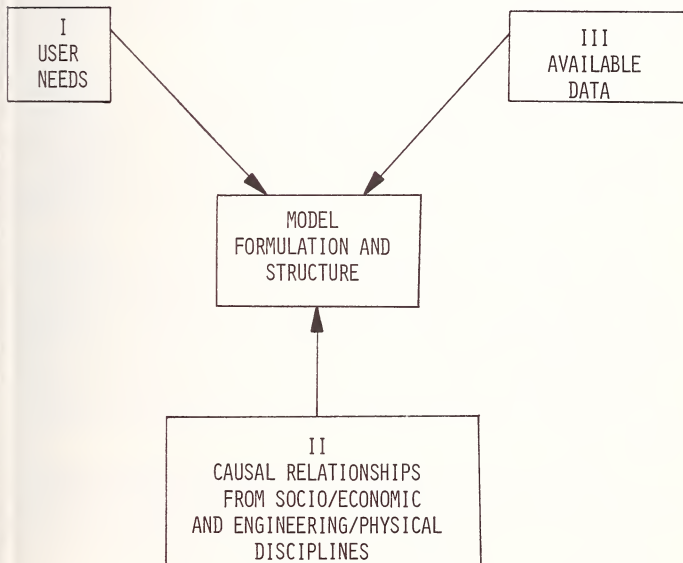
The linkage between the modeler and the causal relationship box represents the incorporation of the results, laws, or "great truths" about the particular problem accumulated from previous work; such information is usually very discipline-oriented, not problem or needs oriented. Finally, the third linkage between the modeler and available data acts as a key element (and usually the most serious constraint) for the model formulation and structuring activity.

The entire process is interconnected. The selected modeling structure must not only be responsive to the user questions but must also be consistent with the available data and known causal relationships. If data are not available, then a different structural approach must be used. Modelers must be careful not to generate a model structure which nicely answers the questions but cannot be supported by existing data. An internal validation task would call for a report confirming that the data requirements generated by the model structure can be achieved. Another internal validation task is to identify new or additional data that will affect model results or structure.

Table 1 lists a set of issues or questions associated with each aspect of the modeling process. The answers to each must be consistent with all of the others. Certainly changing the information desired by user will generally require structural changes in the model which in turn will require data modification and inclusion of different causal relationships. However, such a change in the structure not only requires modified data but so broadens (or sometimes narrows) the range of questions available to the user. Finally, new data permit structural modification which in turn can permit modifications in use.

In the early stages of any modeling effort, the validation function will tend to be concentrated on data aspects. As data requirements are generated, an assessment has to be made not only of the quantity and quality of the available data but also if its form and definition are consistent with the causal relationships of the model. As one looks at the charter of the Energy Information Administration, these activities seem to be the focus of current interest. (DOE, 1977)

As the model proceeds through the development stage and begins to be applied, validation activities occur as a result of interaction with users and other modelers. These validation activities should be incorporated in project funding.

The remainder of this paper is directed to documenting the process of selecting the model and some examples of unanticipated internal validation exercises encountered in industrial energy policy modeling.

TABLE I

INTERNAL VALIDATION ISSUES FOR MODEL STRUCTURE

USE	Does model treat the "right" questions? Are results usable in the decision process? Are the policy variables of importance easily manipulated by the user?
CASUAL RELATIONSHIPS	Is the model formulation consistent with other studies? Are the assumptions reasonable? Are the contraints realistic? If behavioral characteristics are implicit, are the implications understood? Is the level of disaggregation reasonable?
DATA	Do data exist at the level of disaggregation of the model? Are data available (proprietary)? Are data at the "right" level of detail? Are data of reasonable quality?

IV. Background on Choice of Models

At the inception of the industry conservation modeling activity, the first action was to select the kind of model. This choice was determined by user needs. The Division of Industrial Conservation (INDUS) of ERDA* had responsibility for technology-based RD&D programs directed toward improving energy end-use efficiency in industry.** It was immediately recognized that in addition to modeling technologies one had to have models capable of assessing the impact on industrial energy-using capital investment decisions of various price and non-price policies to properly assess their RD&D programs. The merging of ERDA's technology mission and FEA's (now EIA) policy mission in one single agency made such models all the more desirable, since issues of trading off policy options against R&D options are central to DOE's mission

*ERDA, the Energy Research and Development Agency which was absorbed by the newly formed Department of Energy in October 1977.

**Industry is defined quite broadly and includes all energy use outside the residential, commercial, and transportation sectors.

A choice had to be made between an econometric approach and process approach for the structure of the industrial energy policy models. Given the advantages and disadvantages of each as depicted in Table 2, the process approach was preferred. This does not preclude the use of econometric analysis to support and supplement the process models.

Comparison of Process Optimization and Econometric Approaches

First, the user wished to assess the probable impact of introducing specific process technologies into production facilities. The process approach which requires specific representation for each new and existing technology seemed to have a definite advantage over an approach that would (at best) identify the impact of new technologies as some kind of generalized energy efficiency improvement. Moreover, the impact of specific policies (e.g., tax credits), could be assessed with respect to their effect on each technical alternative. An econometric approach would indicate a generalized response to a policy initiative which could not be easily partitioned among the competing alternatives.

Second, the process approach uses more direct engineering information and less econometrically estimated data. Since initially the model users were all engineers, this had the advantage of characterizing the data input and technical alternatives in terms familiar to personnel in the using organization. However, the process models also permit consideration of the policy options of interest to those trained as economists. Since a basic tenet of the BNL approach to energy-economic process modeling is to marry the two methodologies, the BNL staff was equally comfortable with either type of model.

One need of the user is to identify the market penetration of new technologies over an extended time horizon. He should be able to generate such information under varying tax policies, fuel price projections, product demand projections, technology cost assumptions, alternative technology availability, and levels of technology and specific government support. Once again, the explicit process representation permits examination of both the utilization of existing capacity and the change in the composition of capital stock over time.

The response of an industry to changes in energy prices will be related to the age structure of its capital equipment. By using a vintage capital representation for industry, a process model can largely capture this effect. An econometric model captures the effect of vintage stock that existed in the historical period from which the data were obtained. There is no reason to expect, a priori, that this is the same relative vintage and efficiency as in the current period. Even if the data were cross-sectional, there is no reason to believe that the vintaging across geographical regions is either random or uniform. In fact, there is good reason to believe just the opposite.

TABLE 2

MODEL SELECTION

ECONOMETRIC PREFERRED WHEN:	PROCESS OPTIMIZATION PREFERRED WHEN:
Focus on aggregated relationships	Focus on disaggregated relationships, especially investments in and use of specific technologies
Interest in equilibrium	Interest in path to equilibrium
Data limited and aggregated	Disaggregated data available
Behavioral response is unchanged	Behavioral response changes predictably
Institutional structure is constant	Institutions change predictably

NOTE: The models are in reality complements not substitutes; each is capable of answering different (in general) questions, or can be used in tandem to answer the same question.

Economic-Engineering Interface

The models chosen to analyze industrial energy conservation alternatives primarily utilize engineering data. This is an advantage since existing and "near-in" technologies are characterized by factual process descriptive parameters as opposed to statistical estimates. On the other hand, characterization of "down the road" and "over the horizon" technical alternatives are at best characterized with great uncertainty and at worst ignored. The economic assumptions underlying capital investment and output decisions are minimal and explicit, whereas in econometric models they are hidden.

Data

The unavailability of suitable data often acts as a barrier to the use of process approaches. Sparse sets of highly aggregated data forces one into using statistical techniques. These results are often unsatisfactory because the generalized relationships may mask the detailed adjustments taking place within the system. Even worse, the definitional frame is set by the data and often does not exactly correspond to the area being studied. In contrast, industry process models depend upon highly disaggregated data sets.

Fortunately, much disaggregated data exist; but often in the form of single
point estimates. These data were generated, analyzed and improved in an
adversary environment in the early 1970's in connection with the introduction
of environmental requirements on industry and later by the FEA in setting
voluntary industry energy use targets. Since most industrial processes
affected by emissions regulations are high energy consumers, both data sets
were made to order for the industrial energy process models. As a result,
we have a large data base of well-worked data. Data on additional processes
and technical changes will require industry cooperation or further govern-
ment-funded studies.

Although the technical descriptions in the industrial process data
base are quite reliable, we have had considerable difficulty with the cost
data, particularly capital costs. These difficulties arise from several
sources. First, capital costs have risen rapidly and unevenly since studies
were made in the early and mid 1970's and utilizing the relative costs of
that time period may be quite misleading. Second, the definition of capital
costs varies greatly depending upon accounting methods used, treatment of
construction costs, treatment of depreciation, and definition of the boundaries
of the system that is being costed. Third, it is difficult for a process
model to discern between greenfield (new) or roundout (retrofit or plant
expansion) investments; each has a different capital cost for a particular
price of equipment. This is one of the major data weaknesses in the process
models.

V.   Internal Validation Case Histories

The case histories described in this section illustrate validation
issues that have emerged in the development and application of our steel
industry process optimization models. Many of them have not been resolved.
It is this process of identifying the issue and taking appropriate action to
resolve it, followed by careful documentation, that has been defined as in-
ternal validation. Some of the issues described below have not been com-
pletely (or satisfactorily) resolved primarily due to a lack of resources.
Besides identifying and resolving issues, part of the validation process is
to record the unresolved issues and the reasons they remain unresolved.

Case History 1 - How Should We Keep Our Books?

This is an example of a data problem. We found that the capital cost
estimates redundant for the dry coking process in steelmaking vary widely.
Upon examination, it was clear that capital costs quoted by the vendor
were much lower than buyer's estimates of capital cost. (A.D. Little, 1978)
This difference is surprising since the process is used extensively by
steel producers in the USSR. When the question is asked why has it not been
adopted by American industry, the industry says that it is too costly and
the equipment vendors claim the costs are no higher than abroad. In fact,

vendor price is considerably lower than buyer cost because the buyer includes installation, and set up costs and properly charges these to his capital account. He also incorporates in his investment decision the performance uncertainty associated with new technology. Resolution of this issue is important. It may reflect differences in accounting, cost structure, or risk acceptance between the U.S. and U.S.S.R. Is the basis of the problem institutional or accounting and if so, are there policy options available to redress it? We are still looking into this question.

Case History 2 - Where Do You Draw System Boundaries?

This example is similar to the first one in that the issue concerns data. In this case it is technical fuel use data that is in question. The issue concerns the Btu consumption of a blast furnace. One group asserts that on a Btu for Btu basis coke substitution for hydrocarbons in a blast furnace increases total fuel requirements (Tanenbaum, 1977) whereas another group asserts that such substitution decreases energy consumption. (Woolf, 1974) Three reasons were hypothesized to account for this differential.

1. The substitution effects are a function of where the system boundaries are drawn.

2. Substitution of coke for hydrocarbons leading to less Btu use is supported by pilot plant operations whereas the opposite result has been observed in actual operating environments.

3. There are differences between blast furnaces; some may exhibit increased total energy use with hydrocarbon injection, others vice versa.

While this issue remains unresolved at this time, hypothesis one seems the most likely explanation. Our belief is that one group looked only at the impact on blast furnace Btu use, while the other looked at its impact upon the entire steelmaking process.

Case History 3 - How Do You Assess Data From Advocates?

A major problem associated with modeling new technology, particularly when the idea comes from outside the industry, is the difference in technical and economic feasibility postulated by an enthusiastic inventor as contrasted with conservative managers. Our example of this is the oxygen blown blast furnace and coal gasifier which looks extremely attractive when incorporated in the model with cost and performance parameters supplied by the inventor. (Jordon, undated) Industry claims the process will not work as advertised and will not consider adopting it. In this case the model sponsor accepted the industry position, and we have constrained this process out of model solutions for operational purposes. However, the basic process is being

retained as an alternative until we can get further information as to whether the position of industry or the inventor can be accepted. Final resolution will not occur until the inventor can convince someone to build a pilot plant.

Case History 4 – Is Reproducing History Necessary?

In response to accepted validation procedures, the steel model was exercised using product demands, prices, and existing technologies for the post-embargo period 1973-1976 in order to determine how well the model tracked energy consumption during the period. While predicted aggregate energy consumption was within 10% of actual consumption over this period, the behavior of energy intensity (Btu/ton) with respect to capacity utilization in the model was exactly the opposite of that observed during the period (AISI, 1977) Figure 3 illustrates this contradiction.

Actual behavior is explained by the fact that there are large fixed heating requirements in many of the iron and steel making processes which are independent of the level of production over fairly wide ranges of capacity utilization. For example, blast furnaces must be kept hot if any output is anticipated in the near term, because the cost of closing down and then restarting are quite high. This means that reduction in output is accompanied by less than proportional reduction in energy use, which gives rise to the actual behavior illustrated in Figure 3.

Why did the model not reproduce such behavior? The reason is that being a linear programming model, it is mathematically incapable of displaying such scale economies for reasons that need not concern us here. Is this a fatal flaw? Not if one realizes that the model was designed to identify attractive end-using technologies under assumptions of smoothly increasing steel demand without the disruptions caused by business cycles of the sort which produce short-term declines in production. The model was designed for users interested in long run behavior of energy use, not short run response of energy use to business cycles. In this instance, exact reproduction of history is not called for; requiring the model to track history would force a drastic restructuring completely inappropriate for the user's needs. Applying conventional historical validation approaches in this case caused us to use woefully short resources in non-productive ways.

VI. Some Observations

Balance:

Validation is an important activity. Proper performance requires the application of the right kind of resources in the right quantities. Because validation competes with development and application funding, there will be a tendency, in a world where almost all modeling efforts are under-supported, to skimp on validation. The path is indeed a delicate one between

FIGURE 3

Steel Energy Intensity as a Function of Production Level for Mid-1970's
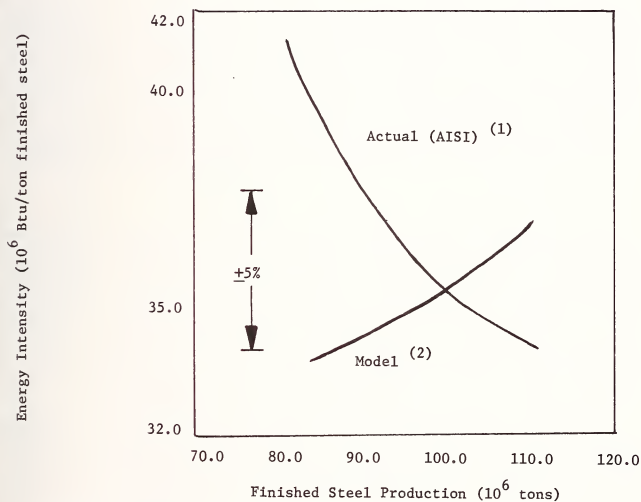
Source: (1) (AISI, 1977)
        (2) Model Runs

a suspect, non-validated, but adequate model and an insufficient, but well validated, analytical tool. Hopefully, the internal validation process would preclude the second outcome by recommending that the modeling effort receive more support or be dropped. The balance between validation and development is precarious and perhaps can best be maintained by recognizing that the purpose of validation is to improve the model.

Overvalidation:

A danger that exists with external validation procedures is that most external validators are also modelers and therefore in competition with organizations whose product they are assessing. This relationship may subtly introduce an unintended bias into the evaluation.

Validation Process:

Too often validation consists of recreating history. Yet, the historical path is made up of the interaction of activities subject to physical laws and economic principles, institutions and behavioral responses. Increased confidence is created by manipulating, examining, and comparing explicit activities with the physical and economic laws that they satisfy. This results in a more fundamental understanding of the structure of the systems and the interaction of the individual parts. By explicitly identifying the behavioral interactions and modifying these where appropriate, enhanced understanding is generated as to how and why the system responds to specific policy alternatives or technology options.

Model Limitations:

All models have design limits or ranges over which they can be applied. These are frequently unspecified. It is easy to fall into the trap of pushing the model beyond its limits without realizing it. One weakness of process models is that they do not incorporate invention, only innovation. They are limited to including only those technologies whose economic and engineering characteristics are quantified in reasonable detail. This could lead to an upward bias in energy use when the models are used in a long-term framework if presently unknown energy conserving technologies were invented and implemented during this period. Part of the validation process should be to identify, document, and publicize model limitations.

Models are not all purpose but are designed to test specific hypotheses and provide certain information. Often to decrease modeling costs and perhaps save time, models are adapted that have been developed for other purposes. When this is done, extreme caution is needed to make certain the model is directed to the user's needs and not to those of some earlier user.

Incorporated in any model are biases often unperceived by the modeler or the user. These unrecognized biases are the most dangerous. When bias is introduced by self-interest it can usually be detected and, more important, taken into account; but when bias results from ignorance, limited background, or from blinders imposed by training or experience, it is very difficult to spot. Special efforts must be made as part of validation activities to identify these unrecognized biases.

Relationship with Industry:

For the modeling program described in this paper, cooperation on the part of industry is imperative. Only they know the full pros and cons of embedding any particular energy using process in their productive facilities. Hidden costs, institutional limitations, and perceptions of risk all enter into the decision at the plant level. These factors will be accounted for as our model development program proceeds. Unfortunately, the climate between government and industry has been increasingly hostile. Industry finds itself pulled in many directions by anti-trust regulations that prohibit information sharing, environmental protection regulations asking for nearly the same information, consumerists who misuse information to show how industry is exploiting the consumer, and environmentalists who may use the information to raise the issue of environmental rape. In this environment, industry has an incentive to remain silent. On the other hand, many industry representatives appreciate government's need for better information in formulating policies to influence industrial energy use. Only time will tell how successfully we can solicit industry support.

VII. Recommendations

1. All models and supporting data bases should specify limits, specific assumptions, critical constraints, and unresolved internal validation issues. The documentation should incorporate the questions the model was designed to address and others that it is capable of addressing. The documentation should also include the applications to which the model is not suitable and the boundaries on data or constraints beyond which the model results can not be interpreted meaningfully.

2. The project plan should incorporate a set of internal validation exercises. These should be carried out concurrently with development and application and should be fully documented. The exercises might include sensitivity analyses on input prices, demands, and process coefficients; alternative formulations for input supply curves or output demand relationships; different levels of process aggregation; and alternative institutional arrangements, among others. In addition, at each stage of development and application new questions will be raised, the answers to which will be provided through internal validation activities. This complete set of planned and ad hoc exercises should be documented and available to users and other interested parties.

3. Validation should be considered as a model improvement activity. Care should be taken that it is not misinterpreted by those using the models to discredit the model. Validation expenditures should be balanced with model development expenditures and must include the cost of internal validation documentation. To use validation budgets most effectively, validation concerns should be rank-ordered and the issues examined from the top of the list down until funds are exhausted. If major issues are still unresolved then there is an indication that the program funds are not balanced. In the process of ordering validation issues care must be taken that the validation actions will not be compromised through incomplete or inaccurate validation data.

4. Internal validation guidelines should be generated and published. While their use may be optional at first, they could become a part of all Federal Government modeling contracts as they are modified and improved through experience.

References

A.D. Little, 1978   A.D. Little, Inc., Research Development and
Demonstration for Energy Conservation: Preliminary
Identification of Opportunities In Iron and Steel
Making (1978).

AISI, 1977   American Iron and Steel Institute, Annual Statistical
Report for 1976, Washington, D.C. (1977).

DOE, 1977   PL 95-91, _Department of Energy Organization Act_,
U.S. Congress, Washington, D.C. (1977).

Gass, 1977   Gass, Saul I., "Evaluation of Complex Models" in
Computer and Operations Review, Vol. 4, Pergaman Press,
Great Britain (1977), pp. 27-35.

Gass, undated   Gass, Saul I., Bibliography: Model Evaluation and
Assessment, University of Maryland, undated.

Greenberger, 1976   Greenberger, Martin, M.A. Crenson, and B.L. Cressey,
Models in the Policy Process, Russell Sage Founda-
tion, New York (1976).

Jordon, undated   Jordon, Robert K., "The Oxygen Blown Blast Furnace
Coal Gasifier", unpublished report.

Marcuse, 1979   Marcuse, William, "Why Should Energy Models Form a
Significant Policy Input In An Uncertain Political
World."

NAS, 1977   National Academy of Science, Applied Research Workshop--
AAAS Meeting, Rensaelervalle Applied Research, June
1976, Center for the Study of Man, Washington, D.C.
(1977).

Tanenbaum, 1977   "Reflections on Steel's Energy Maze" presented at
85th General Meeting of AISI, New York, May 25, 1977.

Woolf, P.L., 1974   "Improved Blast Furnace Operation" paper presented at
Efficient Use of Fuels in the Metallurgical
Industries, Chicago, Illinois, December 9-13, 1974.

MODEL ACCESS AND DOCUMENTATION
Michael L. Shaw
Logistics Management Institute (LMI)
4701 Sangamore Road
Washington, D.C.   20016

Introduction

The objectives of this paper are to describe the needs for the documentation that is necessary to perform evaluation of models and to touch upon the topic of access by outsiders to models in as far as these topics relate to model evaluation. Although our own experience at LMI has been in the documentation and analysis of access issues related solely to energy models these comments should be applicable to most types of models.

Documentation

We must commence by asking the question, "Why is documentation an important topic to discuss at a meeting on model evaluation?" The answer simply is that without documentation it is impossible to evaluate the work that others have done on models and the nature of the documentation will either facilitate or hinder the evaluation process. Documentation is a topic discussed in several of the other presentations in this workshop, and in some of them, formats or requirements for documentation have been proposed. It is documentation that permits outsiders to learn about the model, the input data, and the forecasts that are obtained with the model. The ease with which an outsider is able to use the documentation to obtain the understanding that he or she desires is a reflection of the quality of the documentation.

The credentials with which I discuss this topic are that some three years ago, LMI produced an initial documentation of the PIES Integrating Model. This task has become similar to that of painting a large suspension bridge. As soon as one completes painting the far end, one has to go back to the beginning and start to repaint. This is wholly analogous to the problem of documenting a large and evolving model like PIES which is constantly in a state of change.

I should stress that my insights, are far from unique. In particular, I commend you to a comprehensive, well written and entertaining paper on the subject by Saul Gass (Ref. 1); and also works by House and McLeod and Rafael Ubico (Ref. 2 & 3). In addition, most government agencies have standards for documentation of computer models and computer software including, of course, The National Bureau of Standards (Ref. 4), which Saul Gass describes as being the most comprehensive.

It is important to ask is--"What are the objectives of the documentation, what is it that one is trying to communicate?" The question must be answered by saying that there is no single objective or set of objectives; the specific objectives depend partly on the background of the individual who needs the documentation, partly on the environment, and partly on the use for which the

individual needs the documentation and perhaps the model. As examples of the first of these three factors, mathematicians are clearly interested in precise mathematical statements of the problems that are being modeled; economists are interested in what models represent in an economic sense; regulators might be interested in what the model represents in the way of regulation; and so on. The second and third factors governing the objectives of documentation are the environment in which the user finds oneself and the uses to be made of it. Depending on whether one is a government analyst who is going to have to work with the model; an entrepreneur or businessman whose business environment may be impacted by policy decisions based upon forecast from the model; a policymaker; or the taxpayer who is concerned that the government is using its modeling funds in an appropriate fashion, one's objectives for the documentation will differ.

In consequence I would suggest there exist the following three major objectives for documentation:

- The provision of a description of the model for policy level users which gives them a basis for comprehending its nature and permitting them to perform their own subjective evaluations of its utility to them.

- The documentation should present the model in such a way that is capable of being reviewed and critiqued by other modelers.

- The documentation should provide an archive which permits the progress of changes to the model to be known and permits continuity in the management and development of the model, i.e. it should provide sufficient information to enable new people to construct a functionally identical model.

I suggest that the issues to be answered through documentation are:

- What is the model supposed to do? This should include a specification of the problem or problems that the model is intended to address, the techniques by which the model is intended to operate, the data to be used, and by whom it is intended to be used and operated.

- What does it do? This should describe the applications and users that are actually served by the model as well as the data and modeling techniques actually used. The question, "What does it do?" should permit the potential user to answer the question, "Can I use this model to analyze my specific problem?"

- What doesn't it do? This should describe the limitations of the model and applications for which the model is unsuited.

- How does it do it? This should contain a narrative description of the working of the model, the algorithms in the model, the computer implementation, and other systems or procedures neces-

sary to use the model. Each of these items should in turn contain information at several levels. The narrative description of the model should contain an executive summary and a full narrative description. The narrative descriptions should contain both a statement of the economic representation that the model attempts to match and the policy, regulatory, economic or technical issues that it attempts to address. The section on algorithms should contain both a narrative and a mathematical description of the model. The section on computer implementation should describe the language, the machine on which the model is to be used, the file structure, the job control language, and so forth. The other systems and procedures section should contain information on who initiates a run, who runs the model, who checks its, other resources used in terms of people, cost, and all other related procedures and systems.

- How well does it do it? This should include any information about evaluation of the model.

- What assumptions are made?

- What data are used and whence come they?

- What results or forecasts exist?

- What plans are there for the model?

- What resources does it require to run the model?

- What is the organizational environment in which the model is available?

- What must a potential user do to access the model?

It is worth repeating that each of these issues must be addressed on several different levels.

We can now identify a list of **required documents, they are:**

- A model specification.

- An executive summary description of what the model does.

- A detailed narrative description of what the model does including the principles, structure, and assumptions of the model.

- A complete mathematical statement of the model.

- The computer implementation; perhaps including the model codes.

- A user's guide describing not only how to run the model on a computer but also how to develop scenarios or data or change the model structure to accommodate particular requirements and how to obtain access.

- The data base.

- A description of the validation, verification and audit record associated with the model.

- The future development schedule for the model.

- A record of changes made to the model.

- The results of using the model; both the raw outputs and analyses based upon those results.

This comprehensive list of documents is a goal. It is probably true that no model will ever have all of the appropriate documentation up to date. However, even though this is an ideal or goal, in most instances, the current status of energy model documentation leaves considerable room for improvement. The National Bureau of Standards FIPS Pub 38 (Ref. 4) suggests that the level of documentation required should be linked to the uses that are made of a model and the costs of the modeling effort. This is not something I have pursued here.

In addition to these external documentation needs, there will be internal documentation, not necessarily of primary interest in the model evaluation and assessment process. This documentation would typically include sub-program specifications and internal management matters. Because such documentation may only be of limited value in evaluation, it is not discussed further.

I suggest that the documentation requirements presented here are more comprehensive than suggested elsewhere and seek to illustrate this by the use of Table 1 which shows the approximate correspondence between the 11 types of document proposed by me, the 10 types of document suggested in FIPS Pub 38 (Ref. 4), the 12 types of document identified by Ubico (Ref. 3),[1] and the five documents called for in the recent EIA interim documentation standards (Ref. 5). It seems worth noting that the EIA standards are completely concerned with describing what existing models do. A stronger emphasis on the originally specified goals of the models might be desirable.

It is worth discussing some of these proposed documents in more detail starting with the model specification. Ideally, the specification would be written before any development was done on the model. The model specification should be written to at least two levels. First, there should be the broad requirement of what the model should do and secondly there should be a more detailed specification of the mathematics and the structure the model will have when it is implemented. One issue is who should write a model specification, and my suggestion is that it should be a collaborative effort between the potential user and the modeler. The potential user will express his awareness in one set of terms; the modeler will have a different awareness set, which, in all probability, will constrain the implementation of the user's broad objectives. I do not think that it is inappropriate to point out that the greater the specificity of a specification the better it will be.

---

[1] Of these 12, seven are included in the category 'Programming Documents.'

TABLE 1. CORRESPONDENCE OF DOCUMENTATION REQUIREMENTS

Documents Suggested Here	FIPS PUB 38	SEAS Documentation Type	EIA Standards
Model Specification	Functional Requirements, Data Requirements, System/Subsystem Specification, Data Base Specification	System Definition Document	—
Executive Summary			Model Summary
Detailed Description	Partially Satisfied by User's Manual	Partially Satisfied by Working Papers, Design Papers, Technical Memoranda	Methodology Description
Mathematical Statement			Included in Model Description
Computer Implementation	Partially Satisfied by Program Maintenance Manual, Operations Manual	—	
User's Guide	Partially Satisfied by User's Manual, Operations Manual	Partially Satisfied by User's Guide	User's Guide
Data Base	—	—	Partially Satisfied by Guide to Model Applications
			Included in Model Description
Validation Record	Partially Satisfied by Test Analysis Report	Partially Satisfied by Test Documentation	
Future Development	Partially Satisfied by Test Plan	System Implementation Plan	—
Change Record	—	—	—
Model Results	—	Run Books	—
—	Program Specification	Programming Documents (excluding test documentation and run books)	—

The user's guide is perhaps one of the most important documents. It should identify not only the specific mechanics of how to implement the model on its host computer but also describe how to obtain and validate data, how to model alternate scenario assumptions, how to change the structure of the model, etc. The user's guide should also show how the model is lodged organizationally in its host institution, and should identify the individuals who should be contacted in order to use the model. It should also identify the costs and time delays inherent in using the model.

The validation, verification and audit record should identify those steps that have been taken to perform these functions on the model. It is probable that they would not have been performed in a continuous fashion.

The development schedule for the model should identify future work to be done and should identify milestones, time to reach those milestones, and costs associated with the achievement of those milestones. The development schedule should differentiate between activities which are related to development, test, and implementation.

The publication of results is a significant undertaking. The needs that users have for results will differ according to their environment and perspective. Some users will be interested solely in the conclusions that are based on the analysis of the results while other users will be interested in the detailed numerical values obtained from the model.

The discussion so far has been concerned with physical and written do-cumentation. It should be borne in mind, however, that the purpose of docu-mentation is to instruct potential users and others on the contents of the model. That instruction need not be limited to the written word. Indeed, briefings and demonstrations should be part of the modeler's bag of tools.

It is not immediately obvious who should produce the documentation. Without doubt, the modeler will have the most detailed perception of what goes on in the model. However, it can be argued that the talents of analysts and modelers lie in analyzing or modeling and their talents do not necessarily lie in writing. This is possibly a reflection upon the thought that not everybody can do everything well and that analysts and modelers may be more usefully employed in analyzing and modeling rather than in writing, whereas other people may be more usefully employed in documenting. Further, an outsider coming fresh to the problem and having to produce documentation perhaps as a archivist will have a different perception of what needs to be known about the model to that which the model developer has. Nonetheless, the responsibility for producing the documentation must remain with the modeler.

We should discuss the timing of documentation. If a new model is to be developed, the specification should be written before work on the development of the model commences. Thereafter, the documentation should be produced as the model is developed. Our own experience with PIES has been in performing post facto documentation and, this surely is a much more difficult problem than the parallel documentation which should be produced with a new model.

Clearly, if one has an existing model which is undocumented, one has no option but to perform post facto documentation. This post facto documentation may also include post facto specification of the model. Clearly, if that is to be done, great care must be exercised that it does not specify what the model does do already but what the model was intended to do before its development started. In most of our experience, energy models are evolutionary systems; that is, once developed and used initially, they continue to be developed and improved with time. This evolution also, of course, requires documentation; and, this, too, should be done in parallel with the evolution although again our experience is that if done at all, it is done in a post facto style subsequent to a particular phase of the evolution.

I have no analytical estimate of what it costs to document a model appropriately, i.e., to provide most if not all of the documentation facets described earlier in this paper. Based on our experience of documenting models, I estimate that to do the job well it would take approximately 25 percent of the resources that are consumed in model development and testing. Included in this estimate are writing, typing, organizing, editing, and publishing.

Although the costs of documentation are possibly significant, the costs of not documenting are even more significant in that the sponsor of the model, who is possibly the taxpayer, then forfeits all of the costs associated with the development of the model, because no one can use it except those who have developed it. In addition the sponsor loses the opportunity costs of the development process.

Up till now, we've been talking solely of documentation related to single specific models. However, there are also documentation needs which describe collective bodies of models. In particular, there is a need for a comprehensive catalog or directory of both government and non-government energy models which would permit potential users to identify alternative models which could be used to satisfy their specific requirements. Such a catalog should permit potential users to make the tradeoffs inherent with selection of one model versus another. That is, this catalog should contain uncontroversial descriptions of the different models available. For example, it should identify that a specific model was a natural gas supply forecasting model. It should denote the period for which the forecast was considered to be valid. It should show that the model disaggregated its forecast to a particular regional level, that it was an econometric or some other type of model, that the cost of running the model to obtain a particular forecast was so much, and that, further information on that specific model could be obtained from a particular individual.

Such a catalog would improve the visibility of all models and permit users to narrow the range of their choices when seeking to use a model. In addition, it will aid the identification of modeling needs that are not currently satisfied.

To conclude this section on documentation, let me identify four actions which I believe need to be taken. First, appropriate standards need to be

361

finalized which will be accepted throughout the energy modeling community; and, this should be done through the integration of existing standards together with suggestions developing from this meeting. Secondly, each modeler should explicitly decide upon a documentation philosophy and should consider such aspects as parallel documentation, post facto documentation, in-house documentation, contracted documentation etc. Thirdly, acceptance needs to be made of the budgetry implications; that is, documentation probably is not funded adequately currently if it is to be done to the level of sophistication suggested here. Fourthly, action should be taken to compile and maintain a catalog of energy models of the sort described earlier in this paper.

Model Access

The Energy Conservation and Production Act of August 1976 specifies that the PIES Model will be made accessible to Congressional staff and to the public. We have looked at the what the implementation of this provision means in much more detail than the Act specifies and are pleased that the Energy Information Administration have decided that "access" should be permitted not only to PIES, but also to all other models that EIA possesses. This issue of access is of course related to the issue of evaluation and assessment of models, and what we have done is to specify a number of facets or services that together comprise model access and should permit any interested party with sufficient resources to obtain and evaluate EIA models to the extent that he or she wishes to.

We have proposed that six services be provided, the first being the periodic publication of results or forecasts obtained using the models. This publication should include data used as input to the model. This service has the merits of both being popular with potential users and also being very low in cost. The computer runs will be performed in any case and publication of the output results is very straightforward. It permits potential users to view the scenarios that are being used and possibly, if their scenarios lie between scenarios which have already been run, to interpolate between them.

The second service necessary is the provision of documentation which has been discussed in the first part of this paper.

The third service is the provision of transferable or portable versions of the model. In most cases, for a relatively straightforward model this would not be particularly difficult to implement, but might be difficult to support because of the evolutionary nature of the models. For complex models such as PIES, provision of a transferable version would be a major undertaking.

The fourth service to be provided is that of an "information service", i.e. something in the nature of a bureau service, where people could contact individuals and get the answers to their questions. The drawback to this is that in most cases the people who understand the model and could provide the answers are committed to other tasks in the primary uses of the model. Most people however do find time to undertake some limited activity in this area and contractors are useful in this regard. Services such as these are provided by NASA, DOT, and other government agencies.

The fifth service to be provided would be access to the model on the primary user's computer. This is a difficult thing to arrange and in general we do not feel that this should occur.

The sixth service to be provided is the organization of a "Users Group". The User's Group would provide a forum for the exchange of experiences and opinions about specific models. This practice of having a User's Group is wide spread amongst users of government funded computer and other software and is regarded as a highly beneficial forum for a discussion of results.

References

1. "Computer Model Documentation: A Review and an Approach," Saul I. Gass, University of Maryland, April, 1978

2. Large Scale Models for Policy Evaluation," P. W. House and J. McLeod, John Wiley & Sons, New York, 1977

3. "Documentation Standards for the Strategic Environmental Assessment System (SEAS)," Rafael E. Ubico, Control Data Corporation, August 24, 1973

4. "Guidelines for Documentation of Computer Programs and Automated Data Systems," FIPS PUB 38, National Bureau of Standards, Washington, D. C., February 15, 1976

5. "Interim Model Documentation Standards," Memorandum from George M. Lady, Energy Information Administration, December 4, 1978

Assessment of the READ Model

David Freedman
Statistics Department
University of California, Berkeley

1. Introduction

This paper describes a case study in assessment, focussing on the follow-
ing aspects of the READ model:

- the logic of the equations

- the logic of the fitting procedures

- the quality of the underlying data.

The model, at least in its present form, is judged to be unsatisfactory; some
general conclusions will be drawn at the end of the paper (section 8 below).

READ stands for "Regional Energy Activity and Demography". READ is a
large-scale annual econometric model of the United States, with very fine detail.
The basic geographical unit is the county, and there are almost fifty industries
in the present version of the model. The object is to analyze the impact of
energy policy on regional economic activity.

The model has four basic sectors:

- industrial location (output by industry and county)

- demography (employment by industry and county, labor force and popula-
  tion by county)

- construction (over a hundred building types, by industry and county)

- government (local, state, and federal).

Energy prices appear as explanatory variables in most of the econometric
equations.

There is also a linear-programming transportation sub-model, which in
effect moves industrial output over a transportation network with almost five
hundred nodes. Shadow prices from this sub-model are used as explanatory varia-
bles in the industrial location sector.

Macro-economic variables (GNP and its major components) are treated as exo-
genous, and can be supplied to READ from any of the standard macro-models.
Regional energy prices too are exogenous, and will be supplied to READ from
in-house EIA (Energy Information Administration) models like PIES and its suc-
cessors. In effect, READ disaggregates the macro-level forecasts down to the

county level. Then, it is possible to aggregate back up to any desired geo-
graphical level.

The fitting period for the model is 1965-74, and forecasts will be made out
to 1990.

READ is based on the Maryland model of Harris and Hopkins [1], with further
development work by Hopkins and others, in the

> Office of Applied Analysis
> Energy Information Administration
> Department of Energy
> Washington, D.C.

After the initial development phase, but before the equations were fitted, EIA
decided to review the model. The review process will be discussed in section 2
of this paper. Section 3 will briefly describe the industrial location and demo-
graphic sectors of the model. Section 4 presents a critique of the equations,
and section 5 a critique of the fitting procedures. The data problems are con-
sidered in sections 6 and 7. Conclusions will be found in section 8.

2. The READ review

The review of the READ model was organized by

> George Lady
> Office of Applied Analysis
> Energy Information Administration
> Department of Energy
> Washington, D.C.

and by two of his consultants, Daniel Khazzoom and Richard Ruppert. There were
seven academic reviewers:

> David Brillinger, U.C. Berkeley
> Leon Cooper, Southern Methodist University
> Robert Dorfman, Harvard
> David Freedman, U.C. Berkeley
> Jerry Hausman, M.I.T.
> Karen Polenske, M.I.T.
> Harvey Wagner, University of North Carolina

There was also a meta-reviewer, David Wood (M.I.T.): his function was to review
the review process.

The reviewers were given nearly a thousand pages of documentation on the
READ model, prepared by the modelling group. Then, a meeting was held at which
the model was described and discussed. On the basis of this discussion, George
Lady drew up a list of questions, which were answered in writing by the reviewers.

The READ Model

A second meeting was held to discuss the answers. It became clear that the model had very serious data problems. Indeed, a consensus was reached that

> as specified, and based upon currently available data, the READ model does not justify the resource commitment necessary for its continued development [2].

The issue was then forced: how should EIA do regional impact analysis? A third review meeting was held to consider this question. Three main options were considered:

· Revising READ, for example by increasing the level of geographical aggregation [3]

· Developing some entirely new model

· Dispensing with a model, and doing _ad hoc_ case studies

No consensus was reached. The reviewers did agree on the following point: In a large econometric model with the same fitting period as READ, it would be very difficult to demonstrate the impact of energy prices on industrial location of demographics. The wisdom of developing such a model to do regional impact analysis is therefore questionable.

The main characteristics of the READ review can be summarized as follows:

· Outside reviewers were used.

· The process was reasoned argument, based mainly on documentation supplied by the modelers.

· The reviewers drew on their knowledge of related fields like econometrics, linear programming and statistics.

No attempt was made to compare model forecasts to actual observations, for the following reasons:

· The equations had not yet been fitted, so the model was not in a position to make point forecasts.

· No explicit strategy was available for forecasting exogenous variables.

· There was no valid procedure for estimating the coefficients, or measuring the uncertainty in model forecasts of endogenous variables (section 5 below)

· Most of the endogenous variables in the model are not measured, so there was nothing against which to compare forecasts (section 6 below).

3. A closer look at the READ model

This section will describe the following components of the READ model:

· the industrial location sector

· the demographic sector

· the transportation sub-model

the industrial location sector [4]

"Industry" is a catch-all term. READ industry #1 is "agricultural production," and READ industry #40 is "wholesale and retail trade." The READ industries are defined in terms of the usual Standard Industrial Classification or SIC codes [5]. For instances, READ industry #1 corresponds to SIC codes 01-02, and READ industry #40 corresponds to SIC codes 50-59. The present version of the model has 47 industries, a private household sector, and several government sectors (local, state and federal).

"Industrial location" is a bit of a misnomer. This sector predicts annual changes in the value of output (e.g., sales) by READ industry and county, using a linear regression equation. A "change" is the difference between the value of output for the current year and for the previous year, both measured in 1967 dollars. The independent variables are the following:

· value of output in the previous year

· capital stock

· wage rate in the industry

· land prices

· energy prices

· transportation costs (shadow prices from the transportation sub-model described below)

· tax rates

· macroeconomic variables (GNP, PCE, etc.)

· agglomeration variables (not further specified)

· environmental variables (weather)

The READ Model

the demographic sector [6]

This sector has equations to predict employment, unemployment, wages and population. The first equation in the sector predicts changes in the level of employment in an industry in a county using a linear regression equation. The explanatory variables are:

- level of employment

- change in output

- change in capital stocks, level of capital stocks

- energy prices

- a technology index (not further defined)

The other equations in this sector are similar in conception and will not be discussed here.

the transportation sub-model [7]

In each county, there are two sources of supply for the output of READ's industries, and seven sources of demand.

supply	demand
imports	exports
industrial output	interindustry demand
	personal consumption
	construction
	equipment investment
	state and local government
	federal government

Counties are grouped into the 473 Department of Transportation regions. Supply and demand are summed algebraically over each region, which becomes either a net supplier or a net demander of each type of output.

In effect, a linear program then ships the various industrial outputs over the network, the shipping costs being derived from the Census of Transportation (1967, 1972) and from Interstate Commerce Commission data.

The marginal transportation costs used as independent variables in the READ model are calculated as shadow prices from [this] transportation optimization problem [8].

For a critique of the transportation sub-model, see [9].

4. The logic of the equations

This section will present a series of criticisms of the equations in the READ model. The points will all be straightforward, but should not be brushed aside for that reason. Nor should they be dismissed as "academic." Illogical equations may lead to good forecasts, ones which are validated by events. Indeed, if an equation correctly predicts the future, questions about its logic may seem irrelevant. With READ, however, the argument is almost necessarily a priori. As noted earlier, READ forecasts are very difficult to compare with observations, because READ mainly predicts unobservable quantities. (This is discussed in detail in section 6 below.) In an argument a priori, only sound reasoning compels conviction. On this count, the READ documentation is hardly reassuring: it does not meet the kind of objections raised in this section.

omitted variables

The first point to make is that many important variables are omitted from the equations. For instance, the prices of an industry's major inputs or outputs are not used as explanatory variables in the equations predicting output or employment. Indeed, prices do not appear in the READ data base at all, apart from the land prices, wage rates, and transportation costs mentioned in section 3. (How is the value of output computed? This question will be answered in section 6 below.)

county interactions

The basic geographical unit in READ is the county. But the county is not a natural unit of economic analysis, because county economies interact in complex ways. For example, take the San Francisco-Oakland SMSA [10]. This comprises five counties: Alameda, Contra Costa, Marin, San Francisco and San Mateo. READ industry #40 is "wholesale and retail trade." Changes in the level of employment in this industry for each county is postulated by the model to depend only on other variables describing that same county. However, the reality is very different, because people drive heedlessly across county lines to do their shopping.

To make the point more definitely, suppose the federal government closes an army base in Marin. This exogenous shock will depress wholesale and retail trade in all five counties for some years. In the model, this shock has to be absorbed into the stochastic disturbance terms, because the explanatory variables for wholesale and retail trade in one county do not include army bases in another county. Indeed there are no variables in the equation to capture the impact of one county on another, except the shadow prices from the transportation sub-model which appear in the industrial location equation. (The estimation strategy of READ precludes other choices—section 5 below.) There are no variables to capture the impact of one industry on another, except the "agglomeration variables" in the industrial location equation; nothing is said in the documentation about the strategy for defining or using these variables [11].

Take another case. Santa Clara county (an easy hour's drive south of San Francisco) is one of the great computer manufacturing centers in the United States. Suppose a company there invents a programmable pocket calculator which sweeps the market. This event is a stochastic disturbance to READ industry #26 in Santa Clara county. But employees of this company are going to drive up to the San Francisco-Oakland SMSA to spend their money, a stimulus to wholesale and retail trade in all five counties. This too must be absorbed in the stochastic disturbance terms, because the explanatory variables for wholesale and retail trade in the five San Francisco-Oakland counties do not include computers in Santa Clara.

A third example on county interactions. The University of California is part of READ industry #47 "health, legal and other services." The university has nine campuses, in nine counties. But the rise and fall of employment in the university (and in many other branches of READ industry #47) depends very little on county variables, and very strongly on the state budget. Proposition 13 appears in the stochastic disturbance term for many counties in California, and many years from 1978 onward.

aggregation

There is little interest in predicting the output of one industry in one county, and much aggregation will be done before forecasts are made. So each equation in READ may be suspect, the argument goes, but the errors tend to cancel out during aggregation. That is one defence for modelling at such a fine level of detail.

Cancellation of errors is the _ignis fatuus_ of statistics. If errors are independent, they do tend to cancel--on a relative basis. If the errors are correlated, they pile up, and very fast too: so aggregation can increase the relative error. The errors in READ are likely to be correlated, as noted above. The cancellation argument for modelling in such detail is therefore quite shaky.

coefficients constant across counties

READ uses the same linear functional form for all industries. There is a separate equation for each industry, but in each such equation, the same coefficients are used for all counties. This is illogical. READ industry #47, for instance, is "health, legal and other services," as noted before. In Alameda county, California, this industry almost comes down to U.C. Berkeley. In another county, the main component of this industry might be a private university or a hospital. The industry will show different economic behavior in different counties, and should be described by different equations. The same argument applies to virtually all of READ's industries [12].

an endpoint problem

In the industrial location sector of the READ model, changes in the output of an industry in a county are predicted by a linear equation in which the coefficients are constant across counties; the explanatory variables include GNP and energy prices. This specification presents a serious endpoint problem

for an industry which has been dormant in some county over the fitting period [13].

GNP almost has to come into the equation with a positive coefficient--the same for all counties. Suppose the model starts forecasting, with a scenario for economic expansion. The GNP term on the right hand side of the equation forces positive changes in output. The base is zero, so we get positive output. Dormant industries spring into action all across the United States, driven by the expanding economy. The equation grows wheat in Boston, and strikes oil in Berkeley.

Of course, this economic renaissance could be cut off by an increase in energy prices--for many industries, these must come into the equation with negative coefficients. If the scenario makes OPEC too exigent, the industrial location sector could easily predict negative changes in output. For a dormant industry, the base is zero. Are we ready for the concept of negative output?

The magnitude of these effects is hard to judge. However, economic activity is highly concentrated: in any industry, there will be many, many counties with output either zero or slightly positive over the fitting period. In effect, READ is fitting a straight line to a function which has a definite kink --and then using the line in the vicinity of that kink. For the READ group's comment on this argument, see [14].

explosive autoregressions

Many equations in READ are autoregressive. The employment equation, for instance, has the form

$$L_{jt+1} - L_{jt} = \beta L_{jt} + \text{other terms} .$$

Here, L_{jt} is the level of employment in the industry in county j and year t. Consider forecasting with such an equation, in a scenario where all the explanatory variables (GNP, energy prices, etc.) are constant. Now L_{jt} should converge to its equilibrium value as time goes on. And it would, if the coefficient β were slightly negative. However, the economy was expanding over the fitting period 1965-74, so β is expected to be positive: +0.04 is a typical value in preliminary fits. As a result, L_{jt} can explode geometrically fast to either plus infinity or minus infinity, depending on the coefficients and on the initialization of the other variables. Over a fifteen-year forecasting period, changes in L_{jt} on the order of 50% may be expected--even with all other variables being held constant. This is quite paradoxical. For a more technical discussion, see [15]. The READ group's comment in [14] may be relevant here too.

5. The logic of the fitting procedure

READ uses pooled time-series cross-sectional regressions. To be more definite, consider the employment equation for a READ industry (like wholesale and retail trade). Let L_{jt} be the level of employment in this industry in

county j and year t, and V_{jt} the value of shipments (sales). Let D be the first difference operator: $DX_{jt} = X_{jt+1} - X_{jt}$. The READ employment equation is

(1) $$DL_{jt} = \beta_0 + \beta_1 L_{jt} + \beta_2 DV_{jt} + \beta \cdot U_{jt+1} + \delta_{jt+1}$$

where β_0, β_1 and β_2 are scalar parameters, β is a vector of parameters, U_{jt} is a vector of explanatory variables describing county j in year t, and δ_{jt} is a stochastic disturbance term. Notice that all the parameters do not depend on j or t: they are constant across counties and years, as stated in the previous section.

The READ strategy is to pool the observations from all 3,000-plus counties (indexed by j) and all ten years of the fitting period (indexed by t), fitting equation (1) to the resulting 30,000-plus data points. This pooling is what forces the coefficients in the regressions to be constant across counties. And this pooling obliterates all the economic and demographic inter-relationships between counties.

In its present form, READ uses OLS (ordinary least squares): the parameters are chosen to minimize

(2) $$\sum_{jt}(DL_{jt} - \beta_0 - \beta_1 L_{jt} - \beta_2 DV_{jt} - \beta \cdot U_{jt+1})^2 .$$

The δ's are correlated with each other and with the explanatory variables, so OLS estimates for the parameters and their standard errors are biased.

The READ group proposes to get around this problem by using a variant of two-stage least squares. However, this method cannot be expected to succeed either, essentially for reasons already given. Indeed, the two main assumptions of their proposed method are as follows:

· The stochastic disturbance terms from equations in different sectors are uncorrelated.

· Within a sector, disturbance terms corresponding to different counties are uncorrelated.

But the discussion in section 4 must cast real doubt on these assumptions. If the assumptions are wrong, the estimates for the parameters will still be biased, as will the estimated standard errors. Consequently, significance levels from t-tests on the coefficients are suspect. Since equations are developed by retaining only coefficients which are significant and of the expected sign, even the choice of explanatory variables cannot be justified. These issues are discussed further in [16].

a statistical engineering point

Whatever the δ's are, it is feasible to do the minimization (2) and come up with estimates $\hat{\beta}$. What do these estimates tell us? The crucial point is that economic activity tends to be highly concentrated: for instance, 10% of the counties account for 80% of the manufacturing. As a result, a plot of the

data will show a large cloud of points near the origin for the small counties, with a few points straggling off to represent the big counties. The regression plane will be largely determined by these outliers. What $\hat{\beta}$ does, then, is to measure the contrast between the many small counties and the few big ones. This contrast between the two groups of counties does not seem relevant in pre- dicting how the economies of the counties in either group will evolve over time—or respond to energy policy.

6. The data problem [17]

There is very little solid, relevant county level data (and this situation is unlikely to change in the next few years). As a consequence, most of the data fed into the READ regressions is synthetic: the term of art is "allocated." The accuracy of the allocation procedures is usually impossible to assess, because there is usually no standard of comparison.

This point is crucial to an understanding of the READ model. The alloca- tion procedures will be explained in this section, and then the epistemological status of the main variables in the model will be reviewed. The statistical issues created by allocation will be discussed in section 7.

Probably the single most important variable in the READ model is indus- trial output, which is measured by "value of shipments," e.g., sales. However, value-of-shipments data is collected at the county level only by the Census of Business at five-year intervals (1967, 1972, with 1977 soon to be published). The model needs annual data from 1965 to 1974. This data is derived by an allo- cation procedure which varies a bit from industry to industry. (Interestingly, the Census data is not used, even in the years for which it is available.)

To be perfectly definite, take READ industry #26 (SIC code 35) which manu- factures "machinery, except electrical." Focus on Santa Clara county, California, in the year 1972. County-level payroll data by industry is available from BEA (Bureau of Economic Analysis, Department of Commerce); caveats to this data are discussed below. State-level value of shipments data for manufacturing indus- tries is available from the ASM (Annual Survey of Manufactures, Bureau of the Census). The payroll data is used to allocate the state-level value of ship- ments data to the county level. More precisely, the value of shipments by READ industry #26 in Santa Clara county in 1972 is taken to be the California state value of shipments in the industry in 1972 times the fraction

$$\frac{\text{Santa Clara county payroll in the industry in 1972}}{\text{California state payroll in the industry in 1972}}.$$

The key assumption in this procedure is that the ratio of outputs to labor inputs is constant across counties. This assumption is not plausible. READ industry #26 is quite heterogeneous, so the parts of it in different counties are likely to have quite different labor productivities.

By a charming quirk of fate, "machinery, except electrical" includes both farm machinery and computers [18]. Santa Clara county is known locally as

"silicon gulch." It specializes in computers, not farm machinery, and these two branches of READ industry #26 exhibit very different economic behavior. In 1972, the farm machinery branch had much higher labor productivity (table 1).

Table 1. Labor productivity (value of shipments/payroll) in several
California manufacturing industries in 1972.

SIC Code	Name	Productivity
35	Machinery, except electrical	3.1
352	Farm and garden machinery	4.4
353	Construction machinery	3.8
357	Office and computing machines	2.9

Table 2 below shows what happens if a READ-type allocation procedure is used to create value-of-shipments data for READ industry #26 in twelve California SMSA's: the San Jose SMSA is Santa Clara county. The allocated data is quite good (on a percentage basis) in Los Angeles. It is much less good in San Jose or Fresno, the errors being 10% and 25% respectively. San Jose makes the computers (SIC code 357, with a labor productivity of 2.9). Fresno is in the San Joaquin valley, a rich agricultural area, and specializes in farm machinery (SIC code 352, with a labor productivity of 4.4).

Would more industrial detail help? Probably not. For one thing, three- and even four-digit SIC industries still turn out to be quite heterogeneous. For another thing, the number of firms in each county and industry would get quite small, creating serious instabilities of another kind.

Table 2. READ-type allocated value of shipments for "manufacturing machinery, except electrical" in twelve California SMSA's in 1972 [19]. The recipe is

$$\frac{\text{SMSA payroll}}{\text{state payroll}} \times \text{state value of shipments} .$$

Units: millions of 1972 dollars.

SMSA	Actual	Allocated	Alloc.-Act.
Anaheim	397	403	6
Bakersfield	12	10	- 2
Fresno	66	49	-17
Los Angeles	2,148	2,115	-33
Modesto	6.7	7.2	0.5
Oxnard	47	40	- 7
Riverside	83	79	- 4
Sacramento	28	23	- 5
Salinas	18	15	- 3
San Diego	261	291	30
San Francisco	466	455	-11
San Jose	738	805	72

READ industry #26 was chosen for illustrative purposes only. The alloca-
tion procedure is the same for any manufacturing industry. For non-manufac-
turing industries, the value of shipments by state is unknown, and the BEA pay-
roll data is used to allocate national value of shipments data to the county
level [20]. The inhomogeneity of READ industry #26 is quite typical too: the
economy is just too complicated to divide up into fifty or a hundred homogeneous
"industries."

the BEA payroll data

The BEA payroll figures constitute the key data set in the READ model. The
payrolls are the basis for allocating value of shipments, investments, even
exports and imports. BEA is also the source of the employment levels by
county and industry, the crucial variable in the demographic sector of the
model [21].

A close look at the payroll data is in order. Unfortunately, BEA does not
provide adequate documentation for its procedures [22]. BEA claims that about
80% of the payroll data (by dollars) is based on administrative records--state
unemployment insurance reports. At the county level, however, this claim is
very suspect. Large corporations file only one unemployment insurance report
for each state in which they operate. This report gives the total payroll and
the total employment count for the entire state. It does not give any county-
level detail. The state totals must then be allocated to counties and indus-
trial activities within counties, because corporations often have activities
covered by several different two-digit SIC codes. The allocation is done
either by the state agency or by the BEA, apparently on the basis of the
reported county totals for the small firms which operate only in single coun-
ties. For example, a supermarket chain like Safeway could be distributed
across counties like all the other mom-and-pop grocery stores.

BEA admits that 20% of the payroll data is allocated to counties from
state or national totals. The following sectors of the economy are particu-
larly hard-hit by allocation:

 • Agriculture

 • Transportation

 • Services

 • Government

Agricultural payrolls, for instance, are allocated to counties using shares
from the 1969 Census of Agriculture. Thus, regional differences in the
response of agriculture to increasing energy prices are not captured in the
READ data base. The allocation procedure for transportation should be read in
full [23]. Private household workers, to take one clear example in the service
sector, are distributed to counties according to the 1970 Census of Population.
Local government payrolls are estimated by linear interpolation between the
1967 and 1972 Census of Governments [24]. State government payrolls are esti-
mated from the 1967 Census only, "since the 1972 Census of Governments did not

include the information necessary to prepare a new benchmark." Federal govern-
ment payrolls are allocated from states to counties by year-end employment
headcount, because federal government agencies do not seem to know their payroll
by county.

One thing is very clear. The payroll data, which is used to allocate so
many other variables, is itself the result of a complex allocation process per-
formed at the BEA [25].

a review of the variables

Most of the variables in the READ model are suspect, having been derived
by allocation procedures similar to the ones described above. This applies to
the predicted variables too. As a result, it would be almost impossible to
judge the accuracy of READ forecasts by comparing them to observations: most
of the equations predict variables which are measured only in Census years, if
at all.

The balance of this section will be spent reviewing the main variables of
the model, in rather arbitrary order. The discussion is a bit technical: it
is possible to skip to section 8 without losing the thread of the argument.

Births and deaths. County level data is available from the National
Center for Health Statistics.

Population. Population by county is determined by the Census of Population
every ten years (1960, 1970, ...). The Bureau of the Census and the states
have a cooperative program for estimating county populations in the inter-
census years.

Net migration. READ obtains this for each county by arithmetic from the
birth and death data, and the county population estimates discussed above. In
other words, the Census estimates of net migration are recovered.

Distribution of population by sex, race, and age. This is measured at the
county level only in the Census years. In California, for example, the Census
estimates assume that the distribution does not change between Census years.
Presumably, this assumption has found its way into the READ data base.

Employment and unemployment. The Bureau of Labor Statistics estimates
state-level unemployment rates from the Current Population Survey. READ pro-
rates these to counties, using the 1970 Census of Population ratios of county-
to-state unemployment rates. The assumption is that patterns of unemployment
rates do not change over a decade. For a regional econometric model, this is
unsatisfactory.

The BEA payroll data shows the number of employed persons by county and
industry. (However, as noted above, much of this data results from allocation.)
By summing, READ gets the total number of employed persons in each county. The
number of unemployed persons is then found by solving an equation:

$$\frac{\text{no. unemployed}}{\text{no. employed} + \text{no. unemployed}} = \text{unemployment rate} .$$

Finally, the total labor force in a county is the sum of the employed and unemployed.

Personal income. This is the sum of wages and salaries and non-labor income. Wages and salaries come from the BEA payroll data. The non-labor income was supplied to READ by BEA, but this data is almost entirely the result of allocation--at BEA [26]. For perspective, non-labor income is about one-third of personal income.

Wages and salaries are given by place of work; non-labor income, by place of residence. The two do not mix well; wages and salaries can be adjusted to place of residence only on the basis of ratios observed in the 1970 Census of Population.

In forecasting mode, non-labor income by county is exogenous. It will be very difficult to project non-labor income for 3,000 counties out to the year 1990--especially if we do not know the numbers for any county in any year up to the present.

Energy prices. Coal prices do not seem to be in the READ data base. Electricity and natural gas prices were obtained from the Electric Power Research Institute, by utility district, not by county. They were converted to county level by an unspecified procedure. Fuel oil prices were obtained from the Federal Energy Data System (FEDS) data base, the original source being Platt's Oilgram. Platt's surveys prices in 46 cities. State prices were constructed for FEDS by averaging Platt's estimates for the survey cities in that state. If there were no such cities, the average price for neighboring states was used [27]. Within each state, the READ data base takes the price of fuel oil to be the same for all counties. This is not a good approximation.

The READ documentation suggests that energy availability will be used as an explanatory variable, but there is nothing in the data base to implement this. In forecasting mode, energy prices are exogenous.

Land area. County land areas are reported by the Bureau of the Census.

Land value. The value of farmland is reported by "USDA region" for the years 1966-70, by the Department of Agriculture. There is no indication of how these values are distributed to counties, or estimated for other years. The industrial location sector needs non-agricultural land prices, but these do not seem to be in the data base.

Weather. Temperature and rainfall are reported by the National Oceanic and Atmospheric Administration, and "re-aggregated to a county level." In forecasting mode, weather is exogenous.

Investment. New and replacement equipment investment are handled separately. In manufacturing industries, replacement investment is allocated to states using ASM total investment shares, and then to counties by payroll shares. In non-manufacturing industries, replacement investment is allocated directly to counties by payroll shares. As a consequence, in any particular non-manufacturing

industry, either all counties where the industry operates show positive replacement investment, or all are negative. For a regional model, this is disturbing.

New investment is allocated from national totals to the county level by see construction shares (see below).

Construction. County-level data is available from the F.W. Dodge Corporation. However, documentation on the data-collection procedures of the Dodge Corporation is not available.

Waterborne exports and imports. There is data from the Maritime Commission showing waterborne exports and imports, by industry and county containing the port of embarkation or debarkation.

Landborne exports and imports. There is data from Customs showing the value of landborne exports and imports, by industry and Customs region. Customs regions are subdivided into districts, and there is data giving the total value of landborne exports and imports (aggregated across industries) for each district. The district share of the total for its region is used by READ to allocate exports and imports by industry to the district level. Allocation from districts to counties is done by county payroll share in the transportation industry. (As it happens, this particular payroll series is itself largely the result of an allocation process--at BEA.)

In forecasting mode, exports and imports are exogenous.

Inter-industry demand. Annual data does not exist at the county level. READ creates it from the industrial output by county "data" discussed earlier. Given a figure for the output of an industry in a county, the 1967 BLS national input-output table is used to compute what the demand for intermediate goods "must" have been. This ignores all regional differences in technology. It also ignores all changes since 1967 in relative prices--for instance, the price of labor relative to energy. This is quite unsatisfactory.

Personal consumption expenditures (PCE). PCE is about two-thirds of GNP, and is therefore the largest source of demand for READ's commodities. However, PCE is not measured at the county level. READ creates the data by a very elaborate procedure, whose starting point is the 1972 Census of Retail Trade. This census collected county-level data on sales for ten different types of retail outlets, and SMSA-level data on sales for about two hundred merchandise lines. This data is used to construct county shares of PCE by BEA consumption category. National figures for PCE are then shared down, and converted to county demands for READ's commodities by that 1967 BLS input-output table. However well this worked in 1972, the READ data base ignores any regional trends in consumption: the raw data is available only in 1972.

In forecasting mode, national PCE is exogenous, and is shared down to county level by an econometric equation. The equation, however, is fitted to the "data" just described.

State and local government variables. This sector of the model is quite elaborate, although the documentation is rather sketchy [28]. State and local government activities appear to be endogenous. Variables include expenditures

by function, revenues by type, tax rates and indebtedness. This kind of county-level data is available for local governments from the Census of Governments (1967 and 1972). Also, the Census Bureau has an annual survey of local governments, which provides data for sample counties; however, there is no sensible way of estimating the behavior of individual non-sampled counties. And there is no data on state government expenditures by county, even in Census years. County-level data is needed both for the government sector and for the transportation sub-model. READ creates the data as follows:

> State level controls for state and local government expenditure were obtained for the Morlan model data base at BLS...For the preliminary model, state and local controls will be allocated to counties using [1967 and 1972] Census of Government distributions <u>for</u> <u>local</u> <u>government</u> <u>only</u> [emphasis supplied] [29].

> <u>Federal government</u>. READ almost totally ignores the Federal government, which comes into the data base only through the BEA payroll data. As the documentation notes:

> Data from the Federal Government has not been consolidated to date [30].

7. Statistics of allocated data

Consider the model $y = \beta x + \varepsilon$, where ε is a stochastic disturbance independent of x. The parameter β can be estimated, as usual, by the regression of y on x. Suppose, however, that neither y nor x are directly observable, but are estimated by $\hat{y}$ and $\hat{x}$ respectively. The regression of $\hat{y}$ on $\hat{x}$ may--or may not-- give a good estimate of β. This is the "errors in variables" problem, and it is central to READ--because almost all the data is allocated. This problem will now be discussed in some detail. The analysis will perhaps be oversimplified, but the results are suggestive. The main conclusion is that synthetic data behaves quite differently from real data when regressions are run. In the context of the READ model, this conclusion can be made more specific:

· With 10 years of data and 3,000+ counties, there appear to be 30,000+ degrees of freedom. When the data is allocated, however, this is a gross exaggeration. For example, suppose all the data is derived from national totals using the same allocator. Then the coefficients from a county-level regression based on the allocated data are the same as the coefficients that would be obtained from a regression using the national totals. In fact, there are only 10 degrees of freedom in the allocated data, and the READ-type estimates for standard errors are too optimistic by a factor of $\sqrt{3,000} \approx 50$.

· When different allocators are used for different variables in the equation, as is typical in READ, a substantial bias may be introduced in the estimates, due to the errors in the allocation. Usually, the magnitude of the bias cannot be estimated from the data.

The balance of this section will be spent doing some algebra to justify these conclusions: readers can skip to section 8 without losing the thread of the argument.

The READ Model

Consider first a regression with one explanatory variable and no constant term, say of output on investment. (This cartoon equation does not appear in the READ model, but it illustrates an important point, and keeps the algebra within bounds.) Fix one industry, say "wholesale and retail trade." Let

(1) y_t = national output figure for this industry in year t

(2) x_t = national investment figure for this industry in year t

(3) p_{jt} = payroll in this industry in county j in year t

(4) $f_{jt} = p_{jt}/\Sigma_j\, p_{jt}$.

Thus, f_{jt} is the fraction of the payroll going to county j in year t, and a READ-like allocation procedure [31] is to use

(5) $$\hat{y}_{jt} = f_{jt} y_t$$

(6) $$\hat{x}_{jt} = f_{jt} x_t$$

for "data" on output and investment in county j for year t. The true--but unknown--numbers will be denoted y_{jt} and x_{jt}.

The next step is to choose $\hat{\beta}$ to minimize the sum of squares

(7) $$\Sigma_{jt}(\hat{y}_{jt} - \hat{\beta}\hat{x}_{jt})^2 .$$

Here, j runs over all 3,000+ counties, and t runs over the 10 years in the fitting period 1965-74. Apparently, there are 30,000+ degrees of freedom in (7). However, substitution of (5) and (6) into (7) shows that $\hat{\beta}$ minimizes

(8a) $$\Sigma_t\, w_t(y_t - \beta x_t)^2$$

where

(8b) $$w_t = \Sigma_j\, f_{jt}^2 .$$

In other words, $\hat{\beta}$ is the regression coefficient of the national totals y_t on x_t. The only effect of allocating data to the county level is the introduction of the weights w_t. In principle, these weights are data-dependent, and vary with t. They weight more heavily those years with higher economic concentration. In practice, the w_t's are likely to be essentially constant. In any case, there are only 10 degrees of freedom available for estimating β.

Behind every regression there should be a stochastic model. The kind suggested by READ is

(9) $$y_{jt} = \beta x_{jt} + \epsilon_{jt} ,$$

where y_{jt} and x_{jt} are the true (but unknown) county-level output and investment figures, and ϵ_{jt} is a stochastic disturbance term, assumed to have mean 0. As j and t vary, these disturbances are assumed independent. Furthermore, x_{jt} and

f_{jt} are going to be treated as if they were constant rather than stochastic. These assumptions simplify the analysis, and are similar to those assumed by READ.

Now for a tiny note of realism. There are big counties and little ones, and presumably the likely size of ε_{jt} is bigger in the big counties. It may even be reasonable to suppose that the likely size of the disturbance is approximately proportional to the payroll share f_{jt}, so

$$(10) \qquad \text{var } \varepsilon_{jt} \sim \sigma^2 f_{jt}^2 \; .$$

Under these conditions, it is reasonable to estimate β from a weighted regression, choosing $\hat{\beta}$ to minimize

$$(11) \qquad \sum_{jt} (\hat{y}_{jt} - \hat{\beta}\hat{x}_{jt})^2 / f_{jt}^2 = N \sum_t (y_t - \hat{\beta}x_t)^2$$

where N = 3,000+ is the number of counties. This weighted regression on the synthetic county-level data produces exactly the same results as a regression using the observed national totals, with the spurious 30,000+ degrees of freedom deflated to the corrected 10. This conclusion applies to multiple regression with:

- intercept forced to 0

- the same allocator (f_{jt} in the example) applied to all data

- the error standard deviation estimated proportional to this allocator too.

With a constant term, the algebra is a bit more complicated. The model is

$$(12) \qquad y_{jt} = \alpha + \beta x_{jt} + \varepsilon_{jt} \; ,$$

where the ε_{jt} are stochastic disturbances with mean 0, assumed independent. For OLS, the estimators $\hat{\alpha}$ and $\hat{\beta}$ are chosen to minimize the sum of squares

$$(13) \qquad \sum_{jt} (\hat{y}_{jt} - \hat{\alpha} - \hat{\beta}\hat{x}_{jt})^2 \; .$$

(As before, y_{jt} and x_{jt} are the true but unknown county-level figures; $\hat{y}_{jt}$ and $\hat{x}_{jt}$ represent the allocated data.) Solving the usual normal equations gives

$$(14) \qquad \hat{\beta} = \frac{\sum_t w_t x_t y_t - \frac{T}{N}\bar{x}\bar{y}}{\sum_t w_t x_t^2 - \frac{T}{N}\bar{x}^2}$$

where T = 10 is the number of time periods, N = 3,000+ is the number of counties. The weight w_t was defined by (8b) and exceeds 1/N by Schwarz's inequality. Finally

$$(15) \qquad \bar{x} = \frac{1}{T} \sum_{t=1}^{T} x_t \quad \text{and} \quad \bar{y} = \frac{1}{T} \sum_{t=1}^{T} y_t \; .$$

Again, there really are only 10 degrees of freedom.

The model (12) implies

(16) $y_t = N\alpha + \beta x_t + \varepsilon_t$, where $\varepsilon_t = \sum_j \varepsilon_{jt}$.

Taking x_t and w_t to be nonstochastic, the bias in using synthetic county-level data is easy to work out from (14) and (16);

(17) $$\text{bias} = E(\hat{\beta}) - \beta = \alpha \frac{\sum_t w_t x_t - \frac{T}{N}\bar{x}}{\sum_t w_t x_t^2 - \frac{T}{N}\bar{x}^2} .$$

Since there are only a few large counties, w_t will be appreciably larger than $1/N$, so the bias in (17) can be considerable.

An alternative is to start from (16), using the observed national data instead of the synthetic county-level data: this eliminates bias, and the reduction in degrees of freedom is more apparent than real.

Of course, the model (12) can be revised to incorporate the idea that a small county should have a small intercept and small error. Suppose, as is compatible with the READ allocation scheme, that the intercept is proportional to the payroll share f_{jt}, and so is the likely size of the error:

(18) $y_{jt} = \alpha f_{jt} + \beta x_{jt} + \varepsilon_{jt}$, var $\varepsilon_{jt} \sim \sigma^2 f_{jt}^2$.

Then, α and β would be chosen to minimize the weighted sum of squares

(19) $\sum_{jt} (\hat{y}_{jt} - \hat{\alpha} f_{jt} - \hat{\beta} \hat{x}_{jt})^2 / f_{jt}^2 = N \sum_t (y_t - \hat{\alpha} - \hat{\beta} x_t)^2$.

Again, the weighted regression is equivalent to discarding the synthetic county-level data, and running the regression on the observed national-level aggregates. This conclusion applies to multiple regression when

· the same allocator is used for all variables

· the intercept is proportional to the allocator

· the error SD is proportional to the allocator.

Using the national aggregates directly would have two advantages:

· less machine time gets burned up

· the degrees of freedom get properly deflated to 10.

These conclusions do not apply when different allocators are used for different variables. In such cases, however, substantial bias may be anticipated, because we have an "errors in variables" situation. As an illustrative case, consider another cartoon regression of output on investment, where output is allocated from state totals, while investment is allocated from national totals. The new point is the interplay between the two different geographical levels.

Section 7. Statistics of allocated data

Following previous notation, let

(20) y_{st} = output in the industry in state s in year t.

Abbreviate s(j) for the state containing county j, and let

(21a) $$h_{st} = \sum_{i \in s} f_{it}$$

(21b) $$g_{jt} = f_{jt}/h_{s(j)t} \; .$$

Thus, h_{st} is the state s share of the national payroll, and g_{jt} is the county j share of the state payroll.

In this situation, the county-level output figure may be allocated as

(22) $$\hat{y}_{jt} = g_{jt} y_{s(j)t} \; .$$

However, the county-level investment figure is still given by (6). The model is still (9); the intercept is forced to 0.

Two cases have to be considered, according as the regression on the county-level "data" is run weighted or unweighted. Take the unweighted case. Then $\hat{\beta}$ is chosen to minimize the sum of squares

(23a) $$\sum_{jt} (\hat{y}_{jt} - \hat{\beta}\hat{x}_{jt})^2 = \sum_{st} u_{st}(y_{st} - \hat{\beta} h_{st} x_t)^2$$

where

(23b) $$u_{st} = \sum_{j \in s} g_{jt}^2 \; .$$

The index s on the right of (23a) runs over all 50 states, while t runs over the 10 years in the fitting period.

Notice that $h_{st} x_t$ on the right side of (23a) is an allocation of the national investment figure x_t to the state s, by payroll share. The weight u_{st} measures economic concentration. If the payroll for the industry in state s in year t is concentrated in one county, then $u_{st} = 1$; otherwise, $u_{st} < 1$. The minimum for u_{st} occurs when the payroll is spread evenly over all counties in the state s. So u_{st} does not seem like a sensible weight. Taking the other case for a moment, it is possible to weight county j's contribution to the sum of squares by $1/f_{jt}^2$; then u_{st} gets replaced by n_s/h_{st}^2, where n_s is the number of counties in state s; this too seems an unreasonable choice of weights for a regression on national aggregates.

Coming back to the unweighted case, (23) is minimized by taking

(24) $$\hat{\beta} = \frac{\sum_{st} u_{st} h_{st} x_t y_{st}}{\sum_{st} u_{st} h_{st}^2 x_t^2} \; .$$

The model (9) implies

(25) $\qquad y_{st} = \beta x_{st} + \varepsilon_{st}$, where $\varepsilon_{st} = \sum_{j \in s} \varepsilon_{jt}$.

Taking payroll and investment to be non-stochastic, the bias in $\hat{\beta}$ can in theory be computed from (24) and (25):

(26) $\qquad E(\hat{\beta})/\beta = \dfrac{\sum_{st} u_{st} h_{st} x_t x_{st}}{\sum_{st} u_{st} h_{st}^2 x_t^2}$.

In practice, this expression cannot be estimated, because the state-level investment figure x_{st} is unknown. The "error in variable" here is the discrepancy between the true state-level figure x_{st} and the allocated figure $h_{st} x_t$. The naive "common-sense" approach of putting $h_{st} x_t$ in for x_{st} just assumes the problem away; algebraically, this substitution makes $E(\hat{\beta})/\beta = 1$.

The conclusion is that when different allocators are used for different variables in the equation, as is typical in READ, a bias may be anticipated in the estimates. The magnitude of the bias usually cannot be estimated from the data, but may be large.

So far, the discussion has focussed on ordinary regression, with weights allowed. In principle, instrumental variables can be used to get around the measurement error problem. For READ, this would involve writing down a proper stochastic model, including stochastic equations which link the allocated values to the unobserved county-level data. The source and nature of the stochastic disturbances would have to be analyzed in some detail. Then, it would be necessary to produce some instrumental variables orthogonal to the errors. The orthogonality would have to be argued a priori, on the basis of economic theory, because the errors are unobservable. Omitted variables, county interactions, and other specification errors would all turn up in the disturbance terms, introducing correlations over counties and years. The allocation procedure itself spreads national aggregates (and their errors) over counties, and even over years, another source of correlations. Consequently, neither lagged variables nor national aggregates are plausible candidates for instrumental variables. (For more discussion, see [16].) Therefore, the instrumental-variable approach seems unlikely to solve the errors-in-variables problem. The READ documentation says little of substance about this issue, because the documentation does not address the relationship between the allocated data and the unobserved county-level data.

8. Summary and conclusions

READ represents an ambitious effort to model county-level economic activity, with attention to the impact of energy variables. The modellers have undertaken a difficult assignment, and have worked hard to complete it. However, the prospects of success are remote:

- There is no adequate theory to guide the choice of equations. The READ equations seem ad hoc, mechanical, and in many cases unrealistic.

- The county is not a natural unit of economic analysis, because counties interact with each other in complicated ways.

- There is no rationale for the fitting procedures used in the model.

- There is little solid county-level data. Synthetic data is not an acceptable substitute.

Modellers can sometimes be heard to argue as follows:

It's an imperfect world, we have to do the best we can. Policy-makers want numerical results, so we need a model. The equations may not be perfect, but they're the best we can come up with. We know there are problems with the data, but it's the best we have. If you don't like what we're doing, just tell us how to do it better. Besides, if we don't develop the model, some other agency will, and they'll do it worse.

The result is a variant of Gresham's law, in which bad analysis drives out good, and a fog of misinformation settles over the policy process. With energy statistics these days, who can tell where fact ends and fiction begins?

The READ model is the reductio ad absurdum of the imperfect-world argument. The lesson of READ is that in an imperfect world, making the best model is sometim the wrong strategy. When the basic theory is not well developed or the data is sparse, informal ad hoc analysis by experts may well be superior to a large-scale econometric model. In some cases, it may be even better to tell the policy-maker that his question is unanswerable. This might prompt a search for policies which do not depend on knowing the unknowable.

DISCUSSION

Dr. Greenberg: I want to sort of avoid discussing merits of READ and try to solicit your views on the review process with regard to some specific things that necessarily involve talking about READ, but I am really talking about the review process.

You said that you question some of the functional forms that are in READ. Par-ticularly, you wondered whether the functional form used in one county ought to be the same as the functional form in another county. Given that, did the review process make any attempt whatsoever to suggest what is right?

Dr. Freedman: We did not suggest alternative functional forms to it.

Dr. Greenberg: The second question has to do with the validity on the level of aggregation. Suppose we have two situations, and I would like to ask your view on which you think is the better thing to do. Let's take the extreme situation where there is absolutely no county data of any kind, anywhere, for anything.

Now, under scheme one, you take state data and share down to counties by some reasonable rule of sharing using variables that are available and are at least positively correlated with things that you are sharing down. Then you do the processing, getting transportation, and so on, and then, after you finish, you aggregate back up to state and that gives you the state level output in scheme one.

In scheme two, you start with the state data, you run the process on the state data, and your output is the state data. The question is, which do you think is likely to produce better answers?

Dr. Freedman: That is an issue which was discussed during the review. My own personal inclination would be much more for scheme two.

Dr. Greenberg: You said, I think, and I would like for you to correct me if I state this incorrectly, I think you said, later on, that the goal of READ was to try and find out what is going on in the county and, in approximately one or two sentences later, you said, "And policy makers are really interested in this." If that is what you said, it seems to contradict something said earlier about lack of interest in county data. It also suggests an inconsistency in understanding that, just because the data base and processing is done at the county level, that that necessarily implies that the use of the model is at the county level. The two aren't the same.

Dr. Freedman: Well, it was very hard for us to discover what level of geographic aggregation was intended by the modelers. They do their work at the county level and then they aggregate to some other level, and we could never quite find out from them or from the documentation what they considered the natural level aggregation for READ model forecasts to be. They are set up to do counties. From counties, you can get SMAs or Bureau of Economic Analysis zones, or whatever. So, it wasn't clear to us whether they really planned to forecast at the county level. They tended to deny that at the --

Dr. Greenberg: So, did you presume that the forecast and final output were going to be on county?

Dr. Freedman: The model is set up, and it works that way, it makes its forecast at the county level, and then, when you want to publish a report, you can get in there and aggregate any way you want, but the forecasting is done at the county level.

Dr. Greenberg: Last question. I would challenge your conclusion that suggests that it is possible to increase uncertainty. I would go to the point that, if you remember, that people use models and that the purpose of the model is enlightenment, that the generation of the information in proper hands of analysts that know what all that means can never increase uncertainty, that the addition of information can only reduce or keep uncertainty the same.

Dr. Freedman: Well, I want to say that, after I came back to California from the second meeting, I read a newspaper article which told me how many nuclear reactors California was going to need in the year 2000 and that was attributed to a study in the Department of Commerce, and I presume that that comes out of some kind of model forecasting. So, I think that model forecasts spread well beyond the range of analysts into the general public and, even among analysts. I think people really lose track as to what is data and what is allocated data and what is a forecast and what is not.

So, I really want to stand by the statement that sometimes in some cases models do increase uncertainty.

Dr. Greenberg: Did you say that you knew that this piece of "information" came from a model or are you guessing that--

Dr. Freedman: No, I am guessing. I am guessing. The source was given as the Department of Commerce.

Dr. Greenberg: Okay. You have allocated a statement!

Dr. Freedman: I have allocated a statement. Right, exactly!

Notes

1. C.C. Harris and F. Hopkins, Locational analysis, Lexington Books, 1972.

2. Memorandum by George Lady, December 15, 1978.

3. The READ group proposed modelling at the SMSA level, with a residual non-SMSA region in each state. Some reviewers suggested using the 173 BEA regions, others favored states, still others favored the 10 DOE regions. Some reviewers suggested using many fewer industries, and measuring activity by payroll or employment counts from County Business Patterns (see note 25 below).

   Reviewers felt that increasing the level of geographical aggregation and reducing the amount of industrial detail would improve READ, but not solve its problems.

4. A description of the industrial location sector is in the following briefing document:

   > Estimating a comprehensive county-level forecasting model of the United States--READ, by W.A. Donnelly and others, FEA, 1976.

5. The SIC codes are a United States government standard for classifying economic activity. There are two-, three-, and four-digit codes. The more digits, the finer the classification. For instance:

Code	Description
82	educational services
821	elementary and secondary schools
822	colleges, universities, professional schools and junior colleges
8221	colleges, university and professional schools
8222	junior colleges and technical institutes

   The READ group plans to have about a hundred industries in the final version of the model, including some at the four-digit level.

6. My source is the briefing document:

   > Preliminary version of the structure of the demographic, employment, and income sector of the READ model, by M. Tannen, EIA, Sept. 29, 1979. [The correct year is probably 1978.]

   In the industrial location sector and the demographic sector, county-level predictions may be rescaled to match national control totals: this is reported to me by Frank Hopkins. The impact on the statistical properties of the model has not been analyzed. See note 14 below for more discussion.

7. One source is referenced in note 4. Another is the briefing document:

   > Users Guide to the READ regression file, by N. Gamson and others, EIA TM/EU/78, Dec. 19, 1977. [The correct year is probably 1978.]

8. See p. 38 of the second document cited in note 7.

9. On p. 43 of Harris and Hopkins (note 1 above) it is explained that in the Maryland model, demand is rescaled by a constant factor (across counties) so that total demand equals total supply. Presumably, this is done in READ too; the rationale is not obvious.

    The transportation sub-model was severely criticized by Karen Polenske and other reviewers.

    • The model only covers truck and rail shipments; airborne and waterborne shipments are neglected, and so are pipelines.

    • DOT regions may reflect the nodes in the highway system, but they are not well related to the rail system.

    The transportation model effectively eliminates cross-hauling, and this can severely distort the optimization.

    • All within-zone shipments are eliminated by the netting-out.

    • The netting-out is quite unrealistic, due to the inhomogeneity of READ industries. READ industry #1, for instance, makes both apples and oranges. However, some regions will export apples and import oranges, and these two transactions do not cancel -- for the usual reasons.

    • Each industry's output is handled separately. So it is impossible to consider joint shipping schemes, where e.g. a truck exports some output from READ industry #1 and imports some output from READ industry #2 on the return trip. The reason is that the output of each industry is denominated in dollars, and no conversion factors from dollars to physical units (tons, cubic yards) are available

    The last point is more technical:

    • The solution to a transportation problem is usually degenerate: the dual problem will then have multiple solutions, no one of which gives appropriate shadow prices.

    Some references:

    B.L. Fjeldsted and J.B. South, A note on the multiregional multi-industry forecasting model, University of Utah, 1978.

    D.C. Aucamp and D.I. Steinberg, On the nonequivalence of shadow prices and dual variables, Southern Illinois University, 1978.

    To sum up, the sub-model does not seem to be a good representation of the transportation process. Therefore, it is questionable whether shadow prices from this sub-model are appropriate explanatory variables.

10. An SMSA, or standard metropolitan statistical area, is defined by the Census as an urbanized county, or contiguous group of urbanized counties with strongly inter-related economies.

11. According to pp. 18-19 of Harris and Hopkins (note 1 above), the agglomeration variables are: population density, output of major suppliers, output of major customers.

12. During the review, the modellers decided to put in some regional dummy variables to allow different slopes and intercepts. This mitigates the problem but does not solve it. Another comment is that the theoretical derivation of the employment equation appears wrong. For instance, this derivation starts with a production function which ignores physical inputs, although output is measured by "value of shipments" rather than "value added." And partial derivatives are assumed to be constant over the ten-year fitting period--and the fifteen year forecasting period--so the equation fitted in the model is only a crude linearization of the "theoretical" equation.

13. The same problem comes up for the equation in the demographic sector which predicts changes in the level of employment.

14. The READ group comments that if an industry was dormant in a county over the fitting period, this county is dropped from the data set before fitting the equation. This is a curious procedure. They also write:

    > ...the simulation routine we are developing contains a broad set of initial conditions and checks with regard to applying each equation correctly when forecasting. One of these checks is to ensure that employment changes are not forecast unless industrial production is present. There are additional checks present for alternative situations in which impossible contingencies would otherwise occur.

    No further details are available. However, since the model is more complex than a set of simultaneous linear equations, the rationale for the statistical procedures used to estimate the coefficients is further weakened.

15. In more technical terms, the autoregressions are "unstable" or "explosive." In particular, the usual normal theory approximations do not apply. See

    > T.W. Anderson, On asymptotic distributions of estimates of parameters of stochastic difference equations, _Ann. Math. Statist._ Vol. 30, 1959, pp. 676-687.

16. The READ estimation strategy is outlined in the following paper:

    > A. Havenner and W. Donnelly, Estimation from a pooled time-series of cross-sections of simultaneous equations, FEA, March, 1977.

    This strategy is suggested by

J. Brundy and D. Jorgenson, Efficient estimation of simultaneous equations by instrumental variables, <u>Review of Economics and Statistics</u>, Vol. 53, No. 3, 1971, pp. 207-224.

Also see

J. Hausman, Full information instrumental variable estimation of simultaneous equation systems, <u>Annals of Economic and Social Measurement</u>, Vol. 3, No. 4, 1974, pp. 641-652.

The procedure for developing the equations is discussed in the book by Harris and Hopkins (note 1 above).

Following the notation for equation (1) of section 5, let W_{jt} be the wage rate for the industry in county j and year t. The industrial location equation is

$$(*) \qquad DV_{jt} = \alpha_0 + \alpha_1 V_{jt} + \alpha_2 W_{jt+1} + \alpha \cdot U_{jt+1} + \varepsilon_{jt+1}$$

where ε is the stochastic disturbance term. Suppose there is an exogenous increase over several years in the regional demand for the output of this industry (as in section 4), and suppose the industry adapts over the years in three stages:

· increasing the value of shipments (by volume, or price, or both)

· increasing the wage rate (for instance, by adding overtime)

· increasing the level of employment.

In the beginning, there is an increase in the value of shipments on the left side of the industrial location equation (*) above ($DV_{jt} > 0$); this can be accounted for in our scenario only by making $\varepsilon_{jt} > 0$. Now take the employment equation (1) of section 5. There is no change in employment on the left side ($DL_{jt} = 0$), and a positive change in value of shipments on the right ($DV_{jt} > 0$): so $\delta_{jt+1} < 0$ is needed to cancel the term $\beta_2 DV_{jt}$ and keep the equation in balance. In our scenario, δ_{jt+1} and ε_{jt+1} are negatively correlated, although they come from equations in two different sectors (industrial location, demography). Thus, the first assumption in the READ estimation procedure discussed in section 5 is questionable. The second assumption is even more questionable, for the arguments in section 4 indicate that the disturbance terms for nearby counties are correlated.

Another point: instrumental variables are needed to estimate the coefficients, and it is not clear what instruments are available. Exogenous variables (like GNP) are likely to be correlated with the stochastic disturbance terms. So are lagged endogenous variables. For instance, in the scenario of the previous paragraph, ε_{jt+1} and δ_{jt+1} are both correlated with DV_{jt}. As the adjustment procedures goes on, the correlation spreads: for instance, the ε_{jt} are likely to be autocorrelated over time, so V_{jt} is

The READ Model

correlated with ε_{jt+u}. Likewise, the coefficients in the equations are in effect averaged over counties, and therefore cannot exactly describe the adjustment process in any particular county. The resulting specification error depends on the endogenous variables and turns up in the disturbance term.

A final comment. The READ procedure, as outlined in the Havenner-Donnelly paper, allows explosive autoregressions, in which case the usual asymptotic theory breaks down: see note 17.

17. My source on the READ data base is the second document cited in note 7. Appendix A to that document has a detailed inventory of the data base.

18. This is a feature of the Standard Industrial Classification. Electronic computers were once mechanical, and there is some reluctance to introduce discontinuities by moving companies from one classification to another.

19. The starting point is the 1972 Census of Manufacturers, Vol. III, Area Statistics, Part I, Alabama--Montana. Table 5 in this publication gives state figures for payroll and value of shipments, while table 6 gives the SMSA figures. The first twelve SMSAs in table 6 were used. The allocation is to SMSA rather than county level, because the Census does not publish county-level data. ASM and BEA data were not used. The actual READ allocations, using ASM and BEA data, and going down to the county level, must be worse.

20. The three construction industries (READ 8-9-10, SIC 15-16-17) are exceptions to this rule, as is anthracite mining (READ 3, SIC 11). See pp. 17-18 of the second document cited in note 7.

21. The READ documentation is not entirely clear, and some members of the READ group tell me that employment counts are derived from the BLS Establishment Survey. Of course, this survey covers a giant sample--150,000 establishments. However, when spread across 3,000 counties and 50 industries, this sample melts away, to an average of one establishment per county and industry. Sampling variability just has to kill this stone dead.

22. The best documentation, recommended both by the READ group and by BEA, is the introduction to

> Local Area Personal Income 1970-75. Volume I. Summary. U.S. Department of Commerce, August, 1977, PB 270 880.

By "payroll data" I mean wages and salaries.

23. For example, take railroads.

> County estimates of wages in the railroad industry were based on the biennial employment series for Class 1 railroads (line-haul and switching and terminal companies) prepared by the Association of American Railroads (AAR). These AAR data are for selected

SMSA counties and account for approximately 73 percent of total
railroad employment. For the remaining counties, AAR's residual-
counties employment estimate in each State was disaggregated in
proportion to the railroad employment reported by counties in
the 1970 Census of Population. Estimates for the intervening
years were derived by averaging the biennial benchmark data.
The resulting employment series was used to allocate the State
totals of wages and salaries in the railroad industry. The 1975
AAR data were available for distributing the 1975 State totals.

Source: see note 22.

24. Here is the BEA description of their procedure:

Benchmark estimates of local government wages and salaries were
prepared for each county from the local government payroll data
reported in the 1967 and 1972 Censuses of Government. Estimates
for the intervening years were the products of straight-line
interpolation. The 1973-75 estimates were made in two parts.
For the 372 largest counties, the local government payrolls were
obtained from the annual Bureau of the Census publication, "Local
Government Employment in Selected Metropolitan Areas and Large
Counties." These 372 counties accounted for 71 percent of all
local government wages and salaries. The estimates for the
remaining small counties were made by extrapolating the 1972 bench-
mark estimates by population and using the extrapolated series
to distribute the residual State control totals (derived by sum-
ming the large county payrolls for each State and subtracting
the aggregates from the State control totals).

Source: see note 22.

25. The Bureau of the Census publishes its own payroll data by industry and
county, in County Business Patterns. The data collection procedures are
very well documented and eminently reasonable.

Why doesn't READ use County Business Patterns in preference to BEA?
Three reasons are given:

• County Business Patterns only covers about 80% of the economy.

• County Business Patterns does not provide estimates of non-labor
income; BEA does. Wages and salaries from County Business Patterns
turn out not to mix well with non-labor income from BEA.

• Until 1974, County Business Patterns employment counts were only
for the two-week pay period including March 12.

On the first point, the part of the economy not covered by County
Business Patterns is just the part of the economy where BEA does it all by
allocation. The second point casts further doubt on the BEA non-labor
income figures. The third point is a real limitation on the CBP data.

26. The procedure used by BEA is quite murky. Sample quotes from the documen-
tation (note 22):

That is, the State total for each income component, as taken from the official State income series before adjustment for residence, is allocated to the counties of the State in accordance with each county's proportionate share of the same or some related series that is available on a county basis.

Almost 400 series of separate estimates go into the derivation of the 30 line items shown in the published personal income tables for the SMSA's and counties.

A telephone interview with a senior person at BEA produced the following kinds of responses:

We use the best procedures available.
Do you have any trouble with allocation?
You have to remember how bad the other data is.

27. Federal Energy Data System Analysis and Evaluation, IDEAMATICS, June 30, 1978.

28. Source: see notes 4 and 7.

29. Source: see note 7.

30. Source: see note 7.

31. In READ, "replacement" investment is in fact allocated by payroll shares, but new investment is not (section 6).

A MODELER'S VIEW OF THE READ MODEL ASSESSMENT PROCESS

Frank Hopkins
Energy Information Administration
Department of Energy
Washington, D. C.

I. INTRODUCTION

This paper assesses the recent Regional Energy, Activity and Demographic (READ) model review conducted by a number of distinguished academicians and members of the Office of Analysis Oversight and Access, Energy Information Administration (EIA), Department of Energy (DOE). The paper, authored by the Director of the READ project, is intended to provide suggestions for improving the procedures for reviewing other models, which are expected to be commonplace in the future.

The discussion in this paper presumes that the reader is familiar with the general nature of the Mid-Range Energy Forecasting System (formerly PIES), the regional policy and forecasting problems involved with the interaction between energy and the economy, and the general structure of the READ model.

The remainder of the paper will be divided into seven sections that attempt to chronologically record the development of the review process. Section II will outline the intended role of the READ model in improving the analytical capability of EIA. Section III will discuss the purpose and cost and benefits of the review. The structure of the review will be outlined in Section IV. Section V will present an analysis of the major objections to the county-level READ model as perceived by the Review Committee. The READ model's staff response to modify the model by estimating it at an SMSA level will be discussed in Section VI. The responses of the Committee made at the last review meeting to the SMSA proposal, as perceived by the READ staff, will be summarized in Section VII. The conclusion will present the recommendations of the READ staff on improving future reviews.

II. EIA MID-RANGE MODELING SYSTEM INTERFACE

This section is designed to provide background information to enable the reader to appreciate the larger context of model development in EIA. The role of the READ model in improving EIA's applied analysis capability will be to incorporate feedback from the regional energy projections directly to regional economic projections. In addition, these regional economic projections help to determine regional energy demands through a set of Structural Econometric Energy Demand (SEED) models. Thus, for the first time, the EIA modeling system would embody an energy-economy interaction at the regional level.

This paper is a discussion of the assessment process
as perceived by the author. It is not intended to nor
does it represent a policy statement of the Department
of Energy

Figure 1 will help to clarify this improvement in modeling technique. The two columns in the center of Figure 1 show inputs to the Mid-Range Energy Integrating model, as well as its outputs, and the current form of regional impact analysis. The current analysis flows are shown in solid arrows on the left side of the figure. The expected system of feedbacks using the READ/SEED modeling system is shown in dashed arrows on the right.

Current Analytical System

The current modeling system begins with national macroeconomic projections (DRI). Population, income, and industrial production are shared to regions using fixed regional projections from the Bureau of Economic Analysis in the Department of Commerce (from the OBERS model). These regionalized values are used in the Regional Energy Demand Forecasting model (RDFOR) which constitutes the demand side of the Mid-Range Energy Forecasting System. The Integrating model incorporates regional demands and supplies to produce regional projections of energy consumption and prices. Economic impacts are derived by aggregating selected integrating model results to the national level and inputting them to the macroeconometric model (DRI). Final demands taken from the resulting DRI projections are driven through an input-output matrix inverse to generate estimates of industry output, which in conjunction with a matrix of employment coefficients, creates estimated employment by industry. Regional impacts are obtained by disaggregating these national results using the Regional Earnings Impact System (REIS). Until recently there had been no effort to iterate this system, that is, to feed the DRI results back into the Mid-Range Energy Forecasting System. This reintroduction of second round DRI results has been completed recently, but the results of the regional impact system are not used. Instead, the macroeconomic projections are disaggregated to regions using the same fixed regional shares as were applied in the first round.

The important point to observe is that the current feedback loop loses the regional energy market detail provided by the integrating model since aggregation to the national level must occur before energy results reenter the DRI model. Thus, regional energy conditions cannot affect the regional distribution of economic activity. This is the case with the current regional impacts system whether or not the system is iterated

Proposed READ-SEED System

As shown in Figure 1, regional energy prices from the integrating model are further disaggregated to the state level, using an existing price disaggregation model. These state prices are then used as inputs to READ, thus affecting estimates of county economic activity. The READ

398

FIGURE 1: CURRENT AND PLANNING ENERGY-ECONOMY MODELING INTERACTION
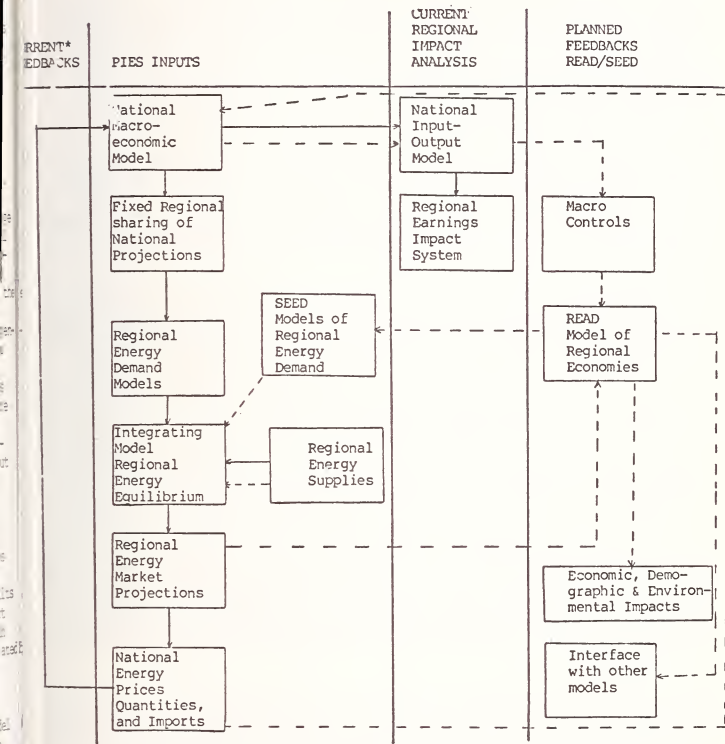
Current relationships are shown by solid line. Planned relationships under READ, SEED System are shown by dashed lines.

399

model is being designed with the option to conform with national con-
trol totals if the user so desires. It is shown operating in this mode
in Figure 1. Thus, integrating model results aggregated to the national
level are fed back into the macroeconometric model (DRI) to obtain
economic impacts at the national level. These results, plus national
impacts by industry from the input-output model, can be used to con-
strain the regional results from READ so that the county estimates
sum to be consistent with national results. In this scenario, READ
provides a regional distribution of activity rather than absolute
levels, but the regional distribution is partially determined by
regional energy markets.

The READ model generates demographic and economic inputs for the SEED
models which generate forecasts at the state level, and being struc-
tural in nature, are well-suited to analyzing the impacts of various
energy policies. This will greatly improve the capability to analyze
the sensitivity of regional energy demands to policy initiatives.

Model Development Plan

The original READ model development plan was divided into three phases,
and had been designed to include an endogenous validation scheme. The
first phase involved construction of a preliminary model which, while
having immediate application in DOE, would also provide: (1) a data
base (2) a software system, and (3) a set of well-specified equations
for use in the second phase of model development. The preliminary model
contains about 350 stochastic equations estimated from county data from
1965 to 1972. Two additional years of data, 1973 and 1974, are being
reserved for post-estimation period simulation tests of the model. The
model was structured into four sectors: (1) industrial location; (2)
population, employment, and income; (3) construction activity; and,
(4) State and local government activity. This preliminary model was
to have been used to provide inputs to the Structural Econometric Energy
Demand (SEED) Models used in the Mid-Range Energy Forecasting System.
While the industrial and construction forecasts of the preliminary model
are developed in considerable detail, the other sectors were to be
estimated in an aggregative form. Thus, this version of the model has
not been designed for use as an extensive demographic or environmental
impact tool.

The second phase of model development would have built upon the results
of the preliminary model. The second generation version, or full scale
READ model, was to have expanded industrial, State and local government,
and demographic, employment, and income sectors. It contained approxi-
mately 800 stochastic equations and will have been estimated using 12

years of county data from 1965 to 1976. The data and software used in the second generation READ model was to have benefited from the validation procedures of the first phase of the model.

Equation estimation in the full model would have also benefited from the preliminary model estimation results. The preliminary model is being estimated using ordinary least squares. While these estimates will not be unbiased, they will generate a set of well-specified equations that can be used as the basis for the full READ model, which was to have been estimated using the iterative two stage least squares procedure developed by Brundy and Jorgenson.[3] This simultaneous equation technique requires a well-specified set of equations before it can be efficiently utilized. In addition to correcting for simultaneous equation bias, the estimation algorithm will also contain corrections for heteroscedasticity and autocorrelation.

The third phase of the READ model development consists of the addition of an environmental sector to the full READ model.

The majority of the criticism by the Review Committee was related to the preliminary version of the model. The substance of their objections is not relevant for the full scale model, as will be illustrated throughout the remainder of the paper.

III. PURPOSE, COST AND BENEFITS

This section will outline the purpose, the financial and staff cost, and the anticipated and realized benefits of the review. The major goals of the review are stated in the introduction to the statement of work provided each member of the Review Committee.

"The purpose of this review is to evaluate the elements of the ongoing model development efforts to determine their essential feasibility and quality and to carefully consider what if any amendments or alternative modeling strategies constitute a pre-ferred means of achieving the overall purposes and objectives of the project."

The READ staff interpreted this statement to imply the review would not be restricted to the model, but would include an analysis of the energy-economy model interface system discussed in Section II. Specifically,

the reviewers were charged with examining the data estimation method-
ology, model component design, validation procedures of data base,
forecasts accuracy, and accuracy of the model applications. In addition,
the reviewers were requested to comment upon the managerial, professional,
and computer service requirements for model implementation.

The costs of the model review can be divided into three areas: review
members consulting fees, EIA staff resources, and model completion
delay. The review members consulting fees were nominal, given the
quality of the reviewers and their own level of effort. Each review
had an initial budget of $7,200, divided into the following categories:
analytical--$5,000, travel--$1,200, and clerical support--$1,000.
The EIA staff resources devoted to the review were more extensive.

The anticipated benefits partially included those items mentioned in the
statement of work. In addition, the READ staff viewed the Review Committee
members as highly qualified, but very inexpensive consultants who could
be used to improve the individual components of the modeling effort. This
was accomplished in a number of areas; particularly, estimation methodology
data development, math programming algorithms and, most importantly, model
development strategy. These areas will be discussed in detail later in
the paper.

The final recommendation for amendments to the READ effort or alternative
modeling strategies to achieve goals of READ were to be the major benefits
of the review to the management of EIA. The complexity of the energy-
economy interface, the paucity of regional energy data, the resource con-
straints of the READ staff, plus the diversity in backgrounds of the
reviewers were the major reasons that a clear concise set of recommenda-
tions were not available at the end of the review process. The divergent
recommendations will be discussed in Section VI.

The READ model was the first model to be reviewed in DOE. There are plans
for evaluating all of the major models in EIA. This process is resource
intensive as indicated by the information in Table 1. While validation
is a necessary concept, careful analysis of the costs and benefits of
each validation study should be undertaken to ensure that each study
involves more than attempting to destroy a model's usefulness through
criticism.

Table 1
Selected Model Validation Expenditures

Applied Analysis	1978	1979
Energy Model Validation		
Procedure Development (NBS)	1 K	248 K
Procedure Development (LASL)	20 K	40 K
Model Validation Symposium (NBS)		25 K
Model Evaluation Program (BNL)	200 K	136 K
Assessment of Coal Supply		
Forecasting System (MIT)	74.9K	25 K

NBS   National Bureau of Standards
LASL Los Alamos National Lab.
BNL   Brookhaven National Lab.
MIT   Massachusetts Institute of Technology
ORNL Oak Ridge National Lab.

IV. REVIEW STRUCTURE

The review process can be disaggregated into 11 chronological sequences: review planning, review contract procurement, READ staff review preparation, first review meeting, reviewers' comments, READ staff response, second review meeting, READ staff response, reviewers' comments, third review meeting, concluding documentation.

The initial review planning was begun in the winter of 1978 at the request of Dr. Lincoln Moses, Administrator, EIA. The original plan was to utilize the National Science Foundation (NSF) to undertake the study in the summer of 1978. Unfortunately, because the formation of DOE resulted in severe disruptions in the contracting process, the NSF could not be engaged to participate in the review. An alternative review procedure was devised whereby the seven independent consultants, with academic backgrounds listed in David Freedman's paper, [11] would conduct the review under the organizational direction of George Lady, Director, Office of Analysis Oversight and Access.

READ Staff Preparation

The work by the READ staff in preparation for the review began in June 1978, and consisted of preparing five documents to be sent to the Review Committee:

403

o The Regional Energy Activity and Demographic (READ) Model:
   Description and Applications (46 pages);

o User's Guide to the READ Regression File (200 pages);

o User's Guide to the READ Estimation and Simulation Software
   (22 pages);

o READ Estimation and Simulation Software Technical Operating Manual
   (73 pages); and,

o READ Model Validation Procedures (51 pages).

The documents, totaling 392 pages, were mailed to the members of the
Committee on September 27, 1978. Additional documents were being pre-
pared for distribution. There activities occupied over 75 percent of
the READ staff from June 1978 to October 1978. The major opportunity
cost of these activities was postponement of completion of the model.
The preliminary version of the READ model is composed of four sectors:
industrial, demographic, construction activity, and State and local
government activity. Initial estimates for the equations in all of the
sectors, except the industrial, were completed during the summer of 1978.
Estimation of the industrial sector was scheduled for completion during
the fall. Testing of the simulation software using these equations
was also initiated during this period. The simulations were designed
to test the robustness of the initial equations, which would be modified
when necessary. The preparation for the review precluded the completion
of this activity. While the reviewers did provide valuable insight into
improving the model structure, the READ staff had speculated that more
directly applicable empirical information would have been obtained
if the simulation tests had been completed. Particularly, since the
READ staff was aware of the problem areas that the Review Committee
addressed during all three meetings.

First Meeting

The first meeting was held on October 12, 1978. Frank Hopkins made
a presentation on the purposes and scope of the READ effort. The
Review Committee members asked questions during the presentation. Their
main themes centered on: (1) the county as a meaningful economic unit;
(2) the appropriateness of OLS in estimating the preliminary model; (3)
the use of synthetic or derived data in the regression equations; and,
(4) the qualifications and background of the READ staff.

The Committee members submitted individual written reports reviewing all
aspects of model development after the first meeting. Discussion of these
reports was the general topic of the second meeting. Their comments were
generally related to judging the scientific integrity of the model against
a perfect standard rather than against alternative analytical procedures.

The panel was composed of highly qualified specialists in econometrics, economics, operations research, and statistics. The comments from members in one area, were in many cases not completely understood by members in other areas. The written comments were general elaborations of the verbal comments discussed in the first meeting. When taken together, the comments appeared to have more serious implications for the model, than if each criticism was analyzed individually.

The READ staff approached the review process from a different perspective than the Committee members. While perfection in model development is a desirable goal, a realistic evaluation must consider the model development effort in relation to resource availability and alternative techniques that may be used to achieve the goals of the model. Thus, the READ staff viewed the major issue as a management of resources to achieve the goals described in Section II. The management of the model has been divided into four areas: data, software, estimation, and simulation procedures.

A paper describing the management plan of the READ effort and the qualifications of the READ staff entitled, "READ Model Management Control Procedures," (RMMCP) was sent to the Committee members before the second meeting.

Second Meeting

The second meeting was held on December 8, 1978. Several review members raised the question of the relevance of forecasting in their written comments prepared for the second meeting. While this is an interesting philosophical question, we did think it was beyond the scope of the purpose of the review. We had the impression the review was concerned with a specific evaluation of the READ model developmental effort. The introduction of this issue raised the much broader question of whether EIA should respond to Congressional requests or inquiries from other offices for analysis of regional policy impacts using models like READ. Since we did not have the opportunity to discuss this issue at the meeting, we mailed a paper by W. Rostow, "Energy, Full Employment, and Regional Development," [20] which contains a concise noneconometric statement of the importance of engaging in the type of regional analysis envisioned for the READ model to the Review Committee members after the meeting. Questions were also raised concerning the legal mandate for EIA to engage in regional analysis. While legislation does mandate that DOE engage in regional analysis, the role of EIA is uncertain.

The Committee members were asked to answer a number of specific questions for the third review meeting by Dr. C. Roger Glassey, Assistant Administrator, Applied Analysis, related to future regional modeling development of EIA. The general nature of the questions concerned the strength of the economy-energy interaction, the need for comprehensive models, and advice on alternatives to READ. There were two specific problems that were to be used as examples in the reviewers' response.

The questions were are follows:

> Question #1: Can the use of energy system variables as explanatory variables for projecting demographic and economic activity variables be dismissed in principle?

> Question #2: What tests should be undertaken to investigate the strength of the energy system/regional demographic and economic activity relationships?

> Question #3: How should EIA proceed to model and project energy system/regional demographic and economic activity relationships?

The problems that were to be discussed in the reviewers response, were how should EIA analyze the socioeconomic and environmental impacts of deregulation of crude oil prices and a moratorium on nuclear power plants.

During the meeting, a number of reviewers proposed that a viable procedure would be to hire a regional energy economist for each region to do the analysis on a case by case basis. Several reviewers and the READ staff objected to this alternative, since the results could not be replicated and may not be consistent with previous studies and data bases.

At the suggestion of Harvey Wagner, the READ staff was given an opportunity to respond in writing to the comments of the Committee before the third meeting. The READ staff response contained a modification in the model development plan and corrections of a number of misconceptions of the reviewers concerning the characteristics of the data base and the model. The modifications included a proposal to estimate the model at a SMSA and remainder of state area rather than the county area.

Third Meeting

The third meeting was held on February 23, 1979. The meeting can be delineated into three phases. First, a presentation of the SMSA proposal. Second, a general discussion of the responses and recommendations of the Committee. The written recommendations were diverse and it was difficult to find a consensus on all points by the Committee. Finally, each reviewer was asked to summarize his position, including modifications of the earlier written statements at the conclusion of the meeting.

The current status of the READ model in EIA has not been fully resolved. The focus of attention has shifted from the technical adequacy of the model to the cost and benefits of developing the model in relation to its projected uses and compared to alternative analytical procedures.

V. MAJOR OBJECTIONS TO THE COUNTY-LEVEL READ MODEL

The major objections to the county-level READ model are primarily valid for the preliminary version of the model. The existence of the majority of the defects of the preliminary model was known to the READ staff during the design phase of the model. But they were not corrected because of resource constraints. The general model development plan called for correcting these deficiencies during completion of the full model. In retrospect, this was an unsatisfactory model development plan.

In summary, the major discussion in the second written reports and meeting can be divided into seven groups: County as a meaningful economic unit, use of derived data, software algorithms, estimation techniques, exogenous variables, forecast validation, and use of the forecasts.

These topics will be discussed in relation to the four READ management task areas: data, software, estimation, and simulation.

Data

The largest proportion of the discussion time was spent on the data and its inadequacy for modeling. The reliability of data was questioned for use in any regional analysis by some members of the panel. This subsection will address these comments in two areas: use of derived data in regional analysis, and corrections of misconceptions of some of the procedures used for deriving the data.

Derived Data - As Karen Polenske stated, "All regional analysts must use allocation methods from time to time for data estimation, although we would prefer not to do so" [18 p. 8]. Ideally, regional economists would prefer to obtain valid data from a secondary source that has been properly documented. Unfortunately, the current Federal data collection and dissemination system does not possess this capability. Thus, larger scale regional modeling projects are forced to use derived data using standardized accounting conventions. The philosophy in designing the data base for the preliminary model was to a large degree conditional upon resource availabilities and the time constrained goal of creating a data base that could be used to demonstrate model feasibility. We were aware that more reasonable procedures for data allocations, as described by Professor Polenske, were available and intended to utilize them in the full model. In retrospect, since organizational and staff constraints (RMMCP) delayed the completion of the preliminary data base until the spring 1978, from its originally planned date of fall 1976, we recommend that this dual managerial data base policy should not be followed in other model development efforts in DOE. It should be noted that the expenditures were not wasted, since DOE has acquired a large amount of actual data at various regional levels and a software system that can be utilized to generate data for use in a restructured READ type model or for use in an alternative analytical system.

While the process of generating derived data creates statistical esti-
mation problems, it also provides the valuable service of alerting
modelers and statistical agencies of the existence of gaps in data used
to construct models at various regional levels. Viewed in this respect
the READ data effort is an unqualified success with respect to county
data. The derived data was developed to replace the data gaps. The
significance of the gaps in the READ data system decline as the data
is presented at higher levels of regional aggregation.

Data Misconceptions - Review of a large modeling project in a short period
of time is a difficult task. Occasionally the reviewer receives a misconcep
tion of subareas of the project through a lack of sufficient documentation
or out-of-date documentation. There were three areas where this occurred
with respect to data definitions which should be corrected: regional detail
of the data, use of export-import data, and derivation of capital stock
data.

At least one reviewer has maintained that, even if more data were avail-
able, it would not be obtained from standardized sources and hence all
regional analysis using data of this type should not be undertaken. We
think that this is a rather extreme and counterproductive view. We feel,
rather, that it is important to properly understand the problems asso-
ciated with using the available data and be cognizant of the potential
biases which these problems can introduce into the model.

The regional detail of at least five data series have been incorrectly
identified by some reviewers. We believe that this incorrect classifi-
cation has contributed to an impression that the data base is weaker than
it actually is and thus is incapable of supporting modeling at any level
of regional detail as David Freedman contends. In her Table 2 classifi-
cation [18] of non-county data, Karen Polenske includes agricultural output,
wages and salaries, weather, transportation flows, but does not include
employment in either Table 1 or 2. The data for agricultural output is
currently available at the county level, although at the time the documen-
tation was written it was allocated from national totals using wages and
salaries. The annual wages and salary data by county have been obtained
from BEA. The quality of the estimation procedure varies by industry.
The data coverage on non-farm establishments has been obtained from the
BLS-UI establishment data file. The wage and salary data for industries
not covered by BLS-UI have been estimated by BEA. The BEA employment
coverage is identical to the wage and salary information.

David Freedman has misinterpreted the wage and salary and employment data
in both his written report and his comments at the second meeting. It
appears that the misconception occurs because he incorrectly uses the

terms wages and salaries and employment interchangeably with regional income and labor force. The latter two are estimated using synthetic techniques as reported in the BEA publication, "Local Area Personal Income," while the former are reported in the BEA county data series described above.

The regional level of the weather data was listed as "not specified" in Table 2 of Karen Polenske's report. The data was obtained from the National Oceanic Atmospheric Administration and is available for 344 climatic zones. While this is not exact county data, the regions have been organized on homogenous weather patterns and thus should provide a good estimate at the county level and a better estimate for SMSA.

The transportation flow data published in the Census of Transportation and used in the earlier Maryland industrial location study [12] was at the national level. Since that time period (1968-1970), the Census has created public use tapes that contain state-to-state estimates of commodity flows. The state-level data, rather than the national as reported in Table 2 is currently used in the READ model. Unfortunately, the documentation on the transportation data has not been completed, so that information was unavailable to the Review Committee.

In reference to the Federal Maritime export and import data, Karen Polenske stated in Table 2 of her report,

"These county data are not provided by appropriate location of production (exports) or consumption (imports) and must therefore be reallocated. Data on foreign transactions are measured at the port of entry. No data are available on the ultimate county for which imports are destined, nor is there information as to the county of origin of exports."

In the second meeting, David Freedman also expressed concern that the Army Corps of Engineers' project to create land export and import data may not be of value to READ, since it is not at the county of origin or destination. The specification of the industrial location sector requires estimates of supply and demand by commodity for use as contraints in the transportation linear programming problem.

Supply is composed of domestic production plus imports at the port of entry. Demand is divided into six sectors: interindustry demand, PCE, equipment investment, construction expenditures, Government expenditures, and exports at the port of embarkation. Thus, the import-export data obtained from the Federal Maritime Commission are at the correct location for use in the READ model.

409

David Freedman [10] utilizes capital stock as a variable in an illustrative example and in his discussion of the quality of the data in his paper for the second meeting. He defines capital stock as directly allocated from national totals using the county share of national payrolls. This is incorrect for the following reasons. There are two measures of capital stock in the model: equipment and structures. He does not differentiate between the two, neither of which is derived according to the procedure he describes. Payroll and output are flow concepts, while capital stock reflects a measure at a specific point in time. Capital stock is estimated each period by adding investment to the previous period's stock and subtracting estimates of depreciation.

The investment in structures estimates are obtained using the county-level Dodge Construction statistics. This data is characterized by structure type, ownership, and selected 2-digit SIC industries. Since the useful life of a structure can vary, but usually exceeds 20 years, the major determinant of the stock in the READ data base will not be the investment, but the initial stock. Initial estimates of capital stock were obtained from various sources.

The estimates of equipment investment which are composed of new induced and replacement investment, in the current preliminary data base are obtained by two different allocative procedures depending upon whether an industry is manufacturing or non-manufacturing. The manufacturing sector allocates state estimates of equipment expenditure for replacement using county to state relative wage and salary shares, while the non-manufacturing sector allocates national equipment expenditures using relative county to national wage and salaries shares. New investment in equipment was not allocated on the basis of wages and salaries, but by using the ratio of county to national construction expenditures for each industrial sector. This procedure makes the implicit assumption that new equipment will be purchased to be used in the new structures when they are constructed. This is very different from assuming, as David Freedman incorrectly believes, that new equipment is a function of last year's wages. The reason that replacement investment is allocated using wage and salary ratios is that this variable will be highly correlated with output and thus can serve as a proxy measure for depreciation.

The development of a data base for the full model would have improved the estimation procedures for capital stock. However since the project did not have unlimited resources for data development, a sub-optimal allocation procedure was used. The current allocation procedure is more reasonable than the system Freedman believes we used. A number of contemplated improvements include wider exploitation of the Census and Survey data to generate estimates of investment at the SMSA level.

This section has reviewed a number of the major criticisms of selected members of the Review Committee. While there have been some individual misconceptions making the data appear weaker than it is, the READ staff

is in agreement that current data in the READ data files are not sufficient to estimate a county-level model. The SMSA and remainder of state areas in our opinion provide the most feasible alternative for three basic reasons: data availability, appropriateness of the economic unit (sales area, elimination of commutation problems, etc), and finally the DOE enabling legislation mandates that analysis be undertaken for SMSA and non-SMSA areas. However, even at the SMSA level, as is standard practice in regional economics, a portion of the data must be derived from other sources. A number of misconceptions concerning the nature of the data have also been outlined. These examples have been used to support our contention that the data is of higher quality than the evaluation of some members of the review committee would suggest. Finally, misconceptions concerning the data, if left uncorrected, could bias the final recommendation of the Committee. In a discussion with David Freedman after the third meeting, he agreed that he had misunderstood the nature of the data as reported in this section, except for the area of the validity of the BEA employment and wage data. He contends that a large portion is actually State rather than county data since firms may fill out only one form when they have several plants in a state.

Software

The READ Software System is a collection of computer programs designed (in part) to provide access to information stored on a computer. The software system is described in [5, 6] and

- o Extracts a subset of required data from over 300 tapes;

- o Compresses the data in a common format onto disk files;

- o Functionally transforms (log, first difference, etc.) or combines (add, divide, etc.) variables in the data base for regression analysis;

- o Executes regression programs with automated audit trails to be used in simulation routines; and,

- o Utilizes simulation routines to generate forecasts.

The ease of use of the system by noncomputer oriented personnel and a modular design which facilitates modification and growth were part of the original design specifications. Consequently, the READ Software System can be reconfigured to operate at the BEA, SMSA, or any other specified geographic area of analysis. In general, the reconfiguration will be transparent to the system user except that the computer processing time will be reduced and the econometric options in the estimation and simulation phase will be increased.

In general, the Review Committee had few negative and several positive
comments on the software system. There were suggestions currently being
implemented by the Committee that will improve the system. A two-period
lag structure could be utilized in the original software. This capability
is being increased at the suggestion of Robert Dorfman and Karen Polenske.
Jerry Hausman's suggestion to identify the actual regional level of the
data (country, SMSA, state, etc.) used in the regressions is being imple-
mented by a simple modification of the variable identification coding
scheme. Currently, the simulation program uses the Gauss-Seidel algorithm
for solving the system of equations. While this is easy to code, it is
subject to convergence problems. Thus, techniques such as dimension
reduction and gradient procedures will be examined for computational
efficiency. In addition, procedures to order the equations into recursive
and simultaneous blocks are being implemented. The SPEAKEASY estimation
preprocessor is being modified to interface with other commercial packages
such as TSP.

Estimation

Planning the estimation of a model is never completed but is constantly
modified as data limitations are recognized, equation specifications are
revised and software constraints are removed by software development.
As with data and software development, we believe the estimation proce-
dures that can be utilized, including revisions suggested by the review
committee, are primarily a function of resources. As Robert Dorfman [9]
notes,

> "The story of specification in large-scale econometric models
> goes like this: In the beginning it is inevitably poor. Hun-
> dreds of relationships have to be specified. To keep the task
> manageable and get the model running at all, it is necessary
> to adopt some stereotypes and impose a fairly rigid, uniform
> specification on some of the classes of equations, for example,
> production functions for different industries or demand functions
> for different classes of consumer goods. Besides, a limited
> staff cannot include experts in all sectors of the economy."

This discussion will be divided into four areas: review of the plans of
the original estimation procedure, proposed revisions in the estimation
plan and specification of the employment and industrial location sectors
of the model.

Initial Estimation Procedures. The RMMCP describes the rationale for
the original estimation plan for the model. The preliminary model
was designed to demonstrate model feasibility and to provide an initial
specification of the equations. This rigid specification system was

to be modified when the full model was estimated. An important input into this modification would be the feedback obtained from state and local officials on the unique regional characteristics that should be included in the estimation of the full model.

Since the original schedule for the preliminary model was to be completed by the winter of 1977 and FEA did not possess a software package capable of handling the large number of county observations, OLS was the only viable alternative at that time. We also realized that the Brundy-Jorgenson technique could have convergence problems. However, its asymptotic equivalence to FIML, [17] its logical computational framework, and use of prior information obtained from the OLS estimates, were the major reasons this estimation procedure was adopted. An attractive feature of the Brundy-Jorgenson technique is that it can be modified to handle pooled cross section and time series estimation problems.

The original documentation neglected to state that dummy variables were being created to test for regional difference in the coefficients of the equations. The preliminary equations in the state and local sector utilized dummy variables to allow regional variation in intercepts. Use of dummy variables to test for differences in the slope coefficients as suggested by a number of the Review Committee members has not begun.

The READ staff is aware of the spurious correlations that could result if a variable that had been used in deriving a dependent variable is used as an independent variable in a regression equation. We have attempted to specify the equations to avoid this problem; however, there may be cases where the individual analyst may have committed this error. The theoretical specification will usually be modified when this problem occurs as the equations are estimated. The internal review and validation processes within the READ staff and Applied Analysis would hopefully be sufficient to uncover these mistakes before the model was complete. Thus, the extensive algebraic "straw man" example, provided by David Freedman [11, section 7] to his review is not operationally relevant for the estimation problem that READ must address. Indeed, as has been mentioned earlier, Freedman uses incorrect definitions of capital stock to achieve his results.

A related but different problem occurs if data that have been synthesized using different procedures are used in an equation. This result, as the READ staff has maintained and as was mentioned by a number of the reviewers, is an "error in variables" problem. The bias created by this problem may or may not be serious relative to the aggregation bias introduced by estimating at higher regional levels with better data, when the coefficients vary over sub-regional units. Consistent estimates could be obtained using instrumental variables. However,

413

concern has been expressed over the reliability of the county level data to provide proper instruments. The severity of this problem declines at higher levels of regional aggregation.

The original estimation plan relied exclusively on OLS and did not include limited dependent variable techniques except for new construction. The rationale for using OLS rather than limited dependent variable techniques in many of the equations was based upon economic rather than statistical theory. The production decision of the firm can be analyzed in short-run or long-run terms. In the short-run, production decisions are only feasible for those areas in which capital stock exists. The long-run decision is primarily concerned with adjustment of capital stock to achieve long run production goals. Thus, the annual change in output or employment equations excluded those regions with zero output or capital stock based upon economic feasibility. Also since the dependent variable is defined as a change it can assume negative values. The specification of the construction equations, however, satisfies the limited dependent variable constraints and should be estimated accordingly. Since we only estimate observations with non-zero output, the number of observations is always substantially below the maximum 3119 per year for these sectors. Thus while there is a loss in degrees of freedom when synthetic data is used as mentioned by the Review Committee, it is not as dramatic as it appears.

Revised Estimation Procedures – The original reasons for using OLS, to complete the preliminary model within a short time-period and because of the lack of software, are no longer valid. The modifications to the software discussed in the previous section and estimation at the SMSA level will permit a less restrictive set of estimation procedures. The use of SPEAKEASY software has not restricted estimation to OLS, but its capability of using iterative estimation forms allows considerable variation in estimation techniques. We propose to follow the suggestions of Jerry Hausman to incorporate regional coefficient differences by estimating the equations separately for each of the 10 Federal regions using FIML procedures. The quality of the data at the SMSA level should support this estimation structure. While the system is already capable of utilizing lags up to two periods, the use of longer lags will be investigated. The increase in the time period for lagged variables may be restricted by the short time series of the data from 1965 to 1974.

Industrial Sector – The specification of the industrial location equations was based upon the earlier work of Hopkins [13] and Harris and Hopkins [12] completed at the University of Maryland. This analysis was undertaken in the late 1960's and was exploratory in nature. The basic structure of the industrial location equations was that changes in output were a function of economic variables including wages, taxes, land values, lagged output and transportation cost and a set of agglomeration variables. The agglomeration variables were designed to reflect a region's attractiveness because of the location of associated input industries or markets.

The transportation variables were actually shadow prices obtained by solving linear programming transportation problems that minimized the cost of shipping an industry's output from sources of supply to demand locations. The Maryland analysis has been reviewed, and the reviews were mixed as indicated by the attachments to Karen Polenske's review. While we agree with the favorable reviews, rather than attack the negative reviews we have attempted to incorporate their valid points in the estimation of the READ model. The criticism of the Maryland model, which does not apply to READ can be grouped into two areas: estimation and forecasting. There are four major criticisms of the estimated equations of the Maryland study: single year, data base, profit motivation as the only factor determining industrial location, mathematical properties of the dual variables of the L.P.

The use of a single year was considered inappropriate for examining changes in industrial location by most reviewers. The current data base for this sector contains data from 1965 to 1975. The 1965-1966 data base was subjected to the same criticism as the READ data base. One of the major difficulties in creating the 1965-1966 data base was the existence of disclosure problems in the County Business Pattern data that was used as the basis for allocating output. The 2-digit SIC wage and salary data, obtained from BEA that is used in the READ model is complete and data is not suppressed because of disclosure problems. Thus, much of the data is proprietary and cannot be released to the public. The use of the SMSA, instead of the county, should improve the data quality, since the amount of derived data used in the estimation will be reduced.

The deriviation of the industrial location equations in the earlier studies was based upon profit maximizing behavior of the individual firm. The deriviation did not assume that the industry would locate to spatially maximize profits, but only that individual firms would increase or decrease their output based upon cost and revenue considerations. The linear programming transportation problem was used to obtain estimates of the cost of transporting a commodity from the production site to its marginal market (defined by highest transportation cost) for supplying industries and the transportation cost of obtaining an input from its marginal supply source for demand regions. The LP utilized the assumption that the systems transportation cost was minimized because individual traders minimized their own transportation cost. Naturally, the shadow prices obtained from the LP are only an approximation to real world marginal transportation cost. If this approximation was close and the theory realistic, these variables would be useful in explaining industrial location. Given the exploratory nature of study, the staff at Maryland was highly encouraged by the early results. The improved data base and extended time period should

be useful in providing additional information on the reliability of using these variables.

While profit maximization may partially explain industrial location, reviewers have stated there are other factors that should be considered including climate and the environment. The READ data base currently has weather data, and the motivation behind specifying an environmental sector was to correct for this deficiency in the earlier study.

The dual variables obtained from the LP are used as shadow prices to approximate marginal transportation cost. A characteristic of large scale transportation problems is that they are degenerate, yielding multiple primal and dual solutions. While this problem was not addressed in the original study, a significant level of research of the READ staff has been devoted to determining the unique properties of the dual variables and their usefulness as shadow prices. Fortunately, the paper by D. Aucamp and D. Steinberg [2] provided as an Appendix in Leon Cooper's review, [4] outlines a solution to the multiple dual solution single shadow price problem.

A criticism of the Maryland study and the preliminary READ model was that only the rail and truck modes were used in the transportation problem. Thus, water carrier and pipelines were excluded. The Army Corps will provide DOE with a tape of distances on the inland waterways so that this mode can be used in estimating transportation variables for the revised SMSA region READ model. Pipelines, while carrying a high volume of shipments by weight, are usually restricted to shipping natural gas and petroleum products. Thus, the omission of this mode should not have a major impact upon the transportation variables of the nonpetroleum industries.

A major computational problem of the Maryland model involved the generation of shadow prices to be used in other forecasts. The transportation code used in the Maryland study, while more efficient than commercially available codes, required 8 minutes of CPU time to solve an average problem and thus was too expensive to be used endogenously within the simulation system. Regression analysis on the historical shadow prices was used as the basis for predicting future values of shadow prices. This procedure introduced biases into the forecasts. Fortunately, the use of the extremely efficient FEANET, [2] eliminates this problem, since the shadow prices will be solved endogenously in the simulation model.

416

<u>Employment Sector</u> – David Freedman has raised several points concerning the specification of the employment equation [21] by Michael Tannen, which he considered "quite representative." He states [10, p. 10]

"The theoretical derivation of this equation is quite inadequate. For instance the derivation starts by positing a production function that ignores physical inputs; it continues by assuming that the partial derivatives of this function are consistent over a 10 year fitting period; and some of the algebra goes away."

We believe that there are three distinct points here, and we would like to respond to them individually. The statement that the production function ignores physical inputs needs clarification, since what we have applied is a rather conventional general production function form relating output to physical capital, labor, and energy. In his footnote 11, Professor Freedman appears to indicate that the omitted physical inputs he is concerned with, are what economists call intermediate goods. This being the case, the point of contention has an easy interpretation. The theoretical production function which we have applied is widely understood to relate to "value added" rather than to "value of shipments." Thus, there is no point of contention with respect to the theoretical specification, but rather with applying this form to the READ data which deals with value of shipments. We think that this is quite a different point, since he raised the issue in the context of model structure. However, we believe that in an empirical context, the point is a valid one. Implicitly, the assumption that is made in using the value of shipments as the measure of output is that there is stable ratio between value of shipments and value added for a particular industry. Initial reaction is that this is plausible, though it does warrant further investigation using Census data.

His next point is that we assume that the partial derivatives of this function are constant over a 10 year fitting period, but we would argue that this is a misunderstanding of the equation specification. We would like to note the obvious point that given a stable production function, the mathematical form of the partial derivatives will be unchanged over time, but the actual value of the partial derivatives will change as factor proportions vary. Perhaps time subscripts should be attached to the variables to clarify this point, but on pages 8-11 of the manuscript <u>Demographic</u> we explained why we were <u>not</u> assuming that the marginal products were invariant over time, but rather offered a way in which such variations could be represented.

The third point raised is a minor one. Professor Freedman states that some of the algebra goes away in forming an estimable equation to relate to the theoretical specification. But the equation in question was never intended to be an accurate algebraic representation. We believed that experienced practitioners would recognize that it was a first-pass linear approximation to a more complicated "true" form.

Simulation and Forecasts

The discussion and written reports on forecasting and simulation were concentrated in four areas: use of forecasts, initial conditions, proper choice of exogenous variables, and accuracy and validation of the forecasts.

Use of the Forecasts - Reference has been made to how well the READ model can forecast the future. Two major points, one relating to uncertainty and the other to statistical bias have been mentioned as sources of forecast error. We feel that it is best to respond to this statement by noting that the READ staff has always recognized that the accurate forecasting of economic and energy-related phenomena, particularly at the regional level, is a difficult task because of the great uncertainties involved. In this regard we have been careful to introduce and distinguish between two major kinds of uncertainty which must affect the validity of forecasts from models like READ. Uncertainty with respect to exogenous events has been traditionally incorporated in the forecasts of the old FEA and the current DOE by the use of scenarios. Scenarios are usually drawn with references to world oil prices, geological find rates and economic growth. The inclusion of scenarios is important in illustrating and delineating the limits of the behavioral relationships represented in the model. But the inclusion of several scenarios also demonstrates that the forecasts should not be construed as predictions of specific variables, but rather are most appropriate in analyzing the impacts of policy or other influences, conditional upon the occurrence of scenario events. We thus feel that in judging the value of READ forecasts, it is extremely important to distinguish one state of affairs in which we do not know everything, from another in which we do not know anything, for quite obviously, the former is usually a far more desirable state of affairs.

The second major type of uncertainty, that of a statistical nature, has also been recognized by the READ staff. We have always maintained that the placement of confidence intervals around the forecasts is the correct presentation of the results, and could be very useful to policy-makers in judging the precisions of the forecasts. But this is also where the second point raised above enters, for clearly such a placement will not alleviate the evils of statistical bias, since if the forecasts suffer from this malady, the confidence intervals will also.

<u>Initial Conditions</u> - David Freedman has claimed to have raised a major problem regarding the use of the regression results in the READ simulation and forecasting routine. Considering the READ employment equations as a representative case, he states in his draft for the second meeting [10, p.10]

> "The READ documentation expects the coefficient of energy price in the equation to be negative: increasing energy prices will cause decreases in the labor force. I find this a bit hard to swallow, but will grant it for the sake of argument. The coefficient is to be the same for all counties. Take a county in which some industry has never operated. So labor force, payroll, value of outputs, and capital stock vanish over the fitting period. Now energy prices statewide indicies rise; the labor force in our industry and county must fall. The baseline employment figure, however, was zero. Is EIA ready for the concept of a negative labor force?"

We feel, however, that this argument is based upon an <u>invalid assumption</u> regarding the specification of initial conditions in the READ model <u>and</u> in <u>all</u> econometrically based simulation models. We first note that the employment equations were not estimated on data for those counties for which there was no industrial production. It was not intended to explain industrial location, but to capture the employment effects of a change in the scale of an existing operation. Thus, it would not be applied to forecasting employment for counties in which positive output is not forecast. But to answer this type of criticism on a more general level, the simulation routine we are developing contains a broad set of initial conditions and checks with regard to applying each equation correctly when forecasting. One of these checks is to ensure that employment changes are not forecast unless industrial production is present. There are additional checks present for alternative situations in which impossible contingencies would otherwise occur.

<u>Exogenous Variables</u> - Proper specifications of initial conditions, while important in preventing improbable forecasts such as negative employment, are also critical in improving the forecasts and in interpreting their usefulness. Examples of the exogenous use of construction forecasts will be used to illustrate these points.

The joint project by TVA, DOL, and DOE in developing the Construction Labor Demand System (CLDS), described more fully in <u>READ Model Interfaces</u>, will provide estimates of construction activity and employment. Many types of construction projects, in particular energy projects,

require an extended time period between initial planning and completion. A goal of the CLDS will be to convert start data, historically obtained from the Dodge Construction data to construction expenditures and employment estimates over the time period of the project. The CLDS will utilize planned energy project data compiled by DOE to initilized their construction expenditure and employment forecasts. It should be noted that energy projects are location specific and also account for 25 - 30 percent of fixed non-residential investments [20]. Subsets of these forecasts are also currently aggregated to DOE regions for use in the DOE Mid-Range Energy Forecasting System (MEFS), formerly PIES. These forecasts will be used as exogenous input to READ to achieve consistency with CLDS and MEFS and to improve the reliability of the READ forecasts. The paper by W. Rostow, indicates the importance of analyzing the level of energy related investment expenditures.

Forecast Validation - Several members of the Committee have indicated that proper forecasts require an understanding of the special conditions which affect each region. In the READ validation procedures we have indicated that we fully intend to engage in an indepth interchange with many local officials regarding our forecasts. Such an interchange can aid us in respecifying the model as development proceeds, and in applying the appropriate constraints on the forecasts for each region.

Most Committee members have indicated that it would be very difficult to check READ forecasts against actual data because of the paucity of real information at the county level. However, by employing SMSA's rather than counties as the basic geographical unit in the estimation procedure, much more real data becomes available. Two techniques for checking the forecasts can then be used. The techniques we had originally envisioned, estimating the model on 1965-74 data and then applying a Theil test on the forecasts for 1975-77 becomes much more meaningful at the SMSA level because of the presence of many more real data series. The technique of backcasting, i.e. seeing how well the forecasts work for previous years, should also be possible to apply in the case of some variables for which sufficient time series exist at the SMSA level.

VI. SMSA READ MODEL

The review process has permitted us to step back from the day to day modeling activity and permitted us to review our original plan of dividing the model development into three phases: preliminary, full, and extension of the full model to include an environmental sector. The primary purpose for developing a preliminary model was to demonstrate feasibility. Once feasibility was obtained, the full model

would be constructed using information obtained from the prototype model. It was anticipated that errors which were detected in constructing the smaller model would be avoided during construction of the larger model. In retrospect, this process contains internal inconsistencies as Jerry Hausman has indicated. Feasibility can only be achieved if the appropriate estimation techniques are utilized so that the sign and magnitude of the regression coefficients will be free of major biases.

In addition, since the preliminary model utilized data that by resource management design is not as high in quality as that in the full scale model, the equations and forecasts would be subjected to justified criticism. Thus, we are abandoning the original model development schedule. We feel that the estimation and simulation of a model at the SMSA level, using improved data and systems estimation procedures and simulation algorithms, answers the most serious criticisms of the current model design.

The remainder of this section reviews the advantages of the currently planned SMSA version of the model in five areas: observational unit, data availability, software modifications, SMSA energy price forecasts, and modified resource management plan.

Observational Unit

The Review Committee was asked to comment on whether the county level geographical detail of READ is necessary to achieve its stated goals in their written report for the second meeting. The responses varied from using individual data, counties, SMSA, BEA, states to the 10 Federal regions.

The major objections to using the county as the observational unit were quality of data and county as a self contained economic area. We had never intended to defend the county observational unit as a self contained economic area, but rather felt that the county data unit was a useful abstraction from which estimation could proceed. The READ Review Committee offered some very persuasive evidence that we had been overly optimistic about estimating the model in this fashion. But in no sense did we feel that the value of the model critically depended upon the county observational unit. The primary advantage of the county unit was the facility it offered in aggregating to other regional areas such as BEA, SMSA, etc. However, we have been convinced that the statistical problems associated with the data and estimation procedures outweigh the value of this aggregation flexibility. The model will still be very useful to DOE, however, if the estimation proceeds at either the SMSA, BEA, or state level, or as suggested by David Brillinger, a combination of regions for which real data exists.

The use of the SMSA region greatly reduces the situs adjustment problem that existed in the county data. Since SMSA consists of contiguous counties of major economically interdependent areas they will incorporate the major retail and labor markets within a region. In addition, the regional differences between places of employment and residence will decline significantly using the SMSA region as an alternative to the county.

Data Availability

Census data provides a large amount of detailed information at 5 year intervals (10 years for housing) that is reported at the county level. It should be noted that there is a disclosure problem with some detailed information at the county level. The annual surveys provide more aggregated data at SMSA, major SMSA or large counties and States. The 2-digit SIC information is adequate for the industrial classification schemes of the preliminary READ model. The original plan was to use row adjustment scheme (RAS) techniques, utilizing the Census county data and the annual SMSA survey data, to derive more accurate county level estimates for the full scale model. In addition, an effort was planned to utilize other data sources, primarily from regulatory agencies to supplement the data. We now propose abandoning this procedure because of estimation problems and data processing resources constraints.

Our new proposal is to directly utilize the SMSA information from the Census surveys of Agriculture, Mining, Manufacturing, Housing and Government, as well as trade association data in the estimation of the model. This procedure has several advantages that should respond to the two major criticisms made during the second review meeting: derived data and appropriate region for analysis. The amount of derived data will be significantly reduced, particularly, output, equipment investment, personal consumption expenditures and the receipts and expenditures of State and local government. The SMSA will also allow the time period of 1967 and 1972 for the State and local government sectors to be expanded to include a time series from 1967 to 1977. There will still be some derived data, but again, this problem always occurs in regional analysis. The data for an SMSA region is also superior to the BEA region, since data is not collected at the BEA region level by any Federal Agency including Census. BEA region data must be aggregated from county data.

Software

Reconfiguration of the system will require no changes in the current estimation portion of the system. The preprocessor creates a temporary file containing all the information necessary to estimate the

coefficients for the specified equations. Under the current design, the user could estimate the coefficients utilizing ordinary least squares or two stage least squares techniques.

This menu was limited due to the amount of information that was to be processed. The reduction in scope to the SMSA level will allow the user to select estimation procedures that were formerly not practical. As in the current system an effort will be made to incorporate off-the-shelf software into the framework to accomplish the computational requirements. For example, SPEAKEASY will continue to act as a data supervisor for the estimation phase. The possible options which can be included are: generalized least squares, three stage least squares, nonlinear least squares, seemingly unrelated regressions and full information maximum likelihood (FIML).

Again the reconfiguration of the software will require no changes in the current simulation code. The specific number of observations per cross-section is a parameter that can be specified by the user. Of course, the processing time will be greatly reduced due to the decrease in information and number of times the entire system of equations must be solved (once for each cross-sectional element, SMSA, etc., for each time-period). The FORTRAN simulation programs utilize the Gauss-Seidel methodology for solving the system of equations. With the reduction in scope, a multitude of commercial simulation procedures can be incorporated into the framework, ranging from dimension reduction to gradient procedures that have historically proven superior to Gauss-Seidel when handling nonlinear relationships and a large system of equations [19].

SMSA Energy Price Forecasts

The preliminary READ model would use State level forecasted energy prices to interface between MEFS and READ. This system would be disaggregated to the SMSA level for the full scale READ model. The State level energy price data has been obtained from FEDS data and consists of aggregated and allocated energy prices. As we have stressed in earlier communications with the READ Review Committee, there are several deficiencies with the price data, particularly with oil prices. Oil prices are collected by major city and allocated to states. A more appropriate estimation scheme would be to use the city or SMSA data directly for the estimation. Energy prices forecasted at the SMSA level would provide more regional variation than state prices, and thus should improve the reliability of READ forecasts and the interface between MEFS, READ and other models.

423

The original resource management plan divided the model development effort into three phases: preliminary, full and addition of an environmental sector. The revised plan eliminates the division between the preliminary and full model. Harvey Wagner suggested that a viable development strategy would be to build a prototype model for 6 to 12 regions, before expanding to a large regional model. We suggested, and he did not oppose, an alternative strategy of reducing the initial size of the model by reducing the number of equations, but keeping the number of regions at the SMSA and remainder of state areas. The equations will be estimated in successive stages as they are required to interface with the SEED models as they are completed. The first set of equations will be used to drive the residential sector, followed by the commercial, transportation, and industrial sectors. This alternative proposal has the advantage of the ease of software and forecast validation implicit in the Wagner purposal, while retaining the cross section time series data base required for efficient estimation.

VII. COMMITTEE'S FINAL RESPONSE

The final response of the Committee at the end of the third meeting to the SMSA proposal, and the requirement for READ type analysis for driving energy demand models and for use in impact analysis, was not a unanimous recommendation to proceed or stop work on the modeling effort. Indeed, the comments ranged from disappointment at abandoning the county model, to the other extreme of David Freedman who advised that the model should be canceled. We have stressed throughout the review process that the primary question was not designing and constructing a perfect model, but the allocation of scarce resources to achieve the goals outlined in Section II. It is also our contention that any viable analytic alternative will have to utilize a data base, software system, estimation routines, and simulation program similar to READ. Thus, continuation of READ cannot be framed as a yes or no question. This point was stressed in the third meeting by the READ staff and apparently began to receive acceptance by the committee by the end of the meeting.

David Brillinger and Leon Cooper stated that the county level model was being abandoned for the SMSA model at too early a stage. They both noted that all the criticism of the derived data were opinions, that should be tested empirically. Brillinger stressed the analogy of the use of derived data in READ, to his previous efforts of projecting election returns for states from key precincts. Brillinger also suggested that the equations should be estimated at the regional level where real data for the dependent variable exists. This would result

in a model that is estimated using a hybrid set of regions including county, SMSA and State. Cooper does not believe that highly aggregative models can answer the major energy policy questions satisfactorily. In reference to READ he,

"...regarded it as a valuable evolutionary tool, to be used to also define data needs and to be modified with new data, studies and results. As such, I continue to regard it as valuable and possibly of potentially greater value than conventional sources of such "projections" and "forecasts".... I see no compelling reasons that a model that has fewer "regions" and fewer variables and which must ultimately be tested in the same way as the READ model, will produce better results. I have every confidence that, in the long run, it will be far more inadequate than READ."

The comments of the other reviewers were not as concise as that of Brillinger, Cooper and Freedman. Robert Dorfman has consistently maintained throughout the review process that

"The relations among major demographic, economic and energy market variables are so intricate that a formalized system of equations is needed to keep track of them and produce a coherent view of what the future may bring forth. Without the discipline of such a set of equations, we are in danger of basing our planning upon impossible contingencies. It is probably unnecessary to add the qualification that sets of equations have no common sense, especially in a world where structural relationships can change without notice, because of legislation or other influences outside the model. Therefore, model forecasts are only one ingredient of an informed judgment, albeit an important one."

In his written response before the third meeting, he submitted an analytical proposal for a modeling system to analyze energy-economy interactions. The general structure of the proposal parallels the MEFS-READ system and its implementation would require a data base, software system, estimation procedures and a simulation routine of the type used in READ. His major concerns were the estimation using OLS, the use of the county as the observational unit, and the theoretical specification of several equations of the model. He stated that the SMSA proposal satisfies ninety percent of the objections that he had with the model.

Jerry Hausman was concerned with the consistency problem between the driving variables used in the current MEFS and the economic impact analysis undertaken using the solution of MEFS. He believed the

major usefulness of READ would be to ensure accounting consistency between the inputs and outputs of MEFS including the incorporation of energy and economy feedbacks. He was concerned, however, that it may be at least two years before high quality energy data is available that could be used to support an integrated MEFS-READ system.

Harvey Wagner's major suggestion on the SMSA model concerned model development strategy, rather than model content. As discussed earlier, we are incorporating his suggestions with modifications to complete the model in stages. He also stressed that noneconometric methodologies may be more appropriate for forecasting the location of large energy construction projects, because of their discrete nature and absence of a time series of events. We are in agreement with this recommendation and plan to use the CLDS system and other analytical techniques for forecasting large discrete projects.

David Freedman indicates in his paper that he does not believe that the data is of high enough quality to warrant further expenditure of resources on development of the model. Karen Polenski withdrew from the review after the second meeting because she had submitted a contract proposal to DOE that could conflict with her participation in the review.

While there was a diversity of opinions on a course of recommended actions for EIA in regional modeling, a major conclusion perceived by the READ staff, and concurred with by C. Roger Glassey, is that the revised SMSA model proposal has satisfactorily answered most of the technical questions raised in the first two review meetings. While the model is not perfect, the data base, software system, estimation procedure, and simulation routines are sufficiently flexible to be utilized in an SMSA READ model or in other viable alternatives that could be used to drive the SEED models and for use in economic impact analysis. The future of the READ effort will depend upon resource availability within Applied Analysis and the relative cost and benefits of READ in comparison to existing alternative models. The READ staff is currently engaged in a cost/benefit survey of existing regional models.

VIII. CONCLUSION

The formal review process on the READ model has been completed. The major benefits of the review were the advice the READ staff received on how to improve the reliability of the model. Unfortunately, a definitive decision has not been reached on whether to proceed with development of the model. Before the review process started the

426

question of whether READ should continue was based on the technical feasibility of the model. The final decision will now be based on resource availability.

The READ staff has formulated several suggestions which we believe should improve the review process. There should be a separation of the Review Committee and the modeling staff and the agency officials financing the review. The presence of agency officials and the modeling staff may have an inhibiting effect on the Review Committee. The agency should not be involved in the organizational procedures of the review, but a chairman elected by the members of Review Committee should be assigned this responsibility. Finally, the Committee should write a final report signed by all members of the Committee with provisions for inclusion of minority opinions. We believe these recommendations will strengthen the review process by making it more impartial and by ensuring that a final report will be delivered to the agency.

References:

1. D. Aucamp and D. Steinberg, "On the Nonequivalence of Shadow Prices and Dual Variables," unpublished manuscript, November, 1978

2. G. Bradley, G. Brown, and G. Graves, "Design and Implementation of Large Scale Primal Transhipment Algorithms," _Management Science_ Vol. 24, September, 1977

3. J. Brundy and D. Jorgenson, "Efficient Estimation of Simultaneous Equations by Instrumental Variables," _Review of Economics and Statistics_, Vol. 53, No. 3, 1971, pp. 207-224

4. L. Cooper, _READ Project Review Phase II_

5. J. Disbrow, M. Durst, F. Hopkins, T. Morlan, and L. Rubin, _User's Guide to the READ Estimation and Simulation Software System_, DOE Technical Memorandum, September, 1978

6. J. Disbrow, M. Durst, F. Hopkins and T. Morlan, _READ Estimation and Simulation Software Technical Operating Manual_, DOE Technical Memorandum, September, 1978

7. J. Disbrow, M. Durst, and F. Hopkins, _READ Model Validation Procedures_, DOE Technical Report, September, 1978

8. N. Gamson, J. Holte, F. Hopkins, B. D. Hong, T. McCallister and T. Morlan, _User's Guide to the READ Regression File_, DOE Technical Report, December, 1977.

9. R. Dorfman, READ Review papers

10. D. Freedman, READ Review Report for the Second Meeting, November, 1978

11. D. Freedman, _Assessment of the READ Model_, January, 1979

12. C. Harris and F. Hopkins, _Locational Analysis: An Interregional Econometric Model of Agriculture, Mining, Manufacturing and Services_, with C. Harris, December, 1972, Heath Lexington Books

13. F. Hopkins, "Transportation Cost and Industrial Location: An Analysis of the Household Furniture Industry," _The Journal of Regional Science_, August, 1972, Vol. 12, No. 2.

14. F. Hopkins and T. Morlan, "The Regional Energy Activity and Demographic (READ) Model: _Description and Application_, DOE Technical Memorandum, August, 1978.

15. F. Hopkins, A. Parkizgari, M. Tannen, READ Interfaces, DOE Technical Memorandum, November, 1978

16. F. Hopkins, READ Model Management Control Procedures, DOE, Technical Memorandum, November, 1978

17. J. Hausman, Full Information Instrumental Variable Estimation of Simultaneous Equation Systems, Annals of Economic and Social Measurement, Vol. 3, No. 4, 1974, pp. 641-652

18. K. Polenski, READ review papers

19. M. Powell and R. Fletcher, "Rapidly Converging Descent Method for Minimization" Computer J. Vol. 6, (1963) p. 163-168

20. W. Rostow, "Energy, Full Employment and Regional Development," presented at the August, 1978 meetings of the American Statistical Association

21. M. Tannen, Preliminary Version: The Structure of the Demographic, Employment and Income Sector of the READ Model, DOE Technical Memorandum, September, 1978

Assessment and Selection
of Models for Energy
and Economic Analysis

Edward A. Hudson

and

Dale W. Jorgenson

Data Resources, Inc.

I. Introduction

This paper is concerned with issues in the selection of models to
be used in analyses of the energy and economy systems. It is directed
towards establishing guidelines for the selection and application of
models. To do this, four existing models are reviewed in terms of their
structure, their strengths, their weaknesses, and their applicability.
From this review emerge guidelines on how to approach the question of
model selection. The models selected for review are illustrative only;
they are intended simply to provide a reasonable coverage of methodology
and applicability from among the many existing models relating to energy
and/or the economy. The four models are:

o the interindustry model used by Clopper Almon and his associates
  at the University of Maryland;

o the Pilot energy model constructed by Dantzig and Parikh at Stanford;

o the Quarterly Econometric Model constructed by Data Resources, Inc.;

o the Hudson-Jorgenson model of energy and economic growth.

These models are being developed and extended continually. We are
basing our survey on 1985: Interindustry Forecasts of the American Economy,
C. Almon et al, D. C. Heath & Co., 1974; Analyzing U. S. Energy Options
Using the Pilot Energy Model, S. C. Parikh, Technical Report SOL76-27,
Stanford University, October 1976; DRI U. S. Model Version 1978C, Data
Resources, Inc., June 1978; The Long Term Interindustry Transactions Model,
E. A. Hudson and D. W. Jorgenson, Federal Preparedness Agency, September
1977. It is possible that more recent versions of some of these models
have different features, e.g. S. C. Parikh has modified Pilot into the
Welfare Equilibrium Model, but for the present discussion these models are
used simply to illustrate issues in model selection, not as a commentary
on the latest version of the models.

II. Almon Model

Clopper Almon and associates have developed a very detailed inter-
industry model for the medium run forecasting of production and spending
patterns. The model has been used to project final demand (consumption,
investment, government purchases and exports) and production levels for
185 industries through 1985. The result is a very detailed set of pro-
jections about the structure of U. S. economic growth over this period.

The structure of the model is based on separate projections of final demand and of input-output coefficients which are brought together to yield estimates of industry outputs. The solution sequence indicates more specifically the structure of the model.

(1) Exogenous data on disposable incomes, prices, government activity and financial variables are introduced.

(2) Econometrically estimated expenditure functions are used to predict consumption and investment final demand on the basis of the exogenous information in (1) and of lagged variables. With government and export expenditure divisions, a total of 131 final demand categories are included.

(3) Allocation coefficients are applied to split each final demand category across the 185 supplying industries. These allocation coefficients are essentially extrapolations of historical expenditure shares.

(4) Input-output coefficients are projected to 1985 by extrapolation, typically using a logistic curve fitted against time.

(5) Final demands for the output of each industry (4) and the input-output coefficients (A) are combined to determine the total output required from each industry $(X = (I-A)^{-1}Y)$.

(6) Labor productivity for each sector is extrapolated and is divided into that sector's output to estimate demand for labor input to that sector.

(7) The sum of labor demands across sectors is related to projected labor supply; if the two are unequal, the estimate of disposable income is adjusted and the solution returns to (1).

(8) Upon solution, the model gives estimates of the 1985 levels of output from each of 185 industries, as well as the inter-industry transactions involving each industry.

Given this model specification and structure, what can be said about the strengths, weaknesses and range of applicability of the model?

The strength of the model lies in the detail of its projections. The output covers a 185 x 131 matrix of final demand expenditures and a 185 x 185 matrix of interindustry transactions. Also, these projections are internally consistent in the interindustry economics sense. There is enough specific information in these forecasts to be of direct interest to some types of business planning such as investment and sales planning. Similarly, this detail may be valuable for government analysts seeking specific information about output and transaction levels for purposes such as environmental impact analysis and pollution projections. In

short, the model provides a bridge for linking macroeconomic variables to detailed forecasts that may be useful for several types of business and government planning.

The weaknesses of the model relate to the key information that must be supplied from outside the model and to the rigidity of economic behavior as represented by the model. Some of the key information - prices, government activity, financial conditions, disposable income - is an input to, rather than being estimated within, the model. This implies that the system is a macro-micro bridge rather than a stand-alone model. Second, the economic behavior contained in the model is completely rigid - the projected allocation of expenditure and inputs relies almost entirely on time trends; consumer and producer responses to influences such as prices are not included nor is the impact of capital, energy or other constraints upon growth. It is unrealistic to expect this constancy and steady behavior to characterize economic growth and structure through 1985.

In view of these strengths and weaknesses, some conclusions can be drawn about the appropriate use of this model. It is not designed for, nor is it suitable for, use concerning economic growth, behavior or technology. It is, however, a good disaggregation framework providing consistency, breadth and some allowance for change in decomposing a given growth path into its industry level implications. The principal value and application of this model is as a bridge between general forecasts and the detailed information needed for many types of analyses in industry and government.

III. Pilot Model

The Pilot model was developed by George Dantzig and Shail Parikh at Stanford. It is an activity analysis model focused on energy supply and conversion but also linked to representations of energy demand and the economy. The model is solved as a multi-period programming problem where total consumption is maximized. The model has been used to make energy projections and analyses over the long term, with the model covering a 40 year period.

The structure of the model involves linear sub-models of energy conversion and economic production which link energy and capital input to consumption and investment final output. The present value of consumption output is maximized subject to energy and capital constraints. Some features of the model are:

(1) Solution is over 40 years, comprising 8 periods of 5 years each.

(2) Economic activity is represented by fixed coefficient (from the 1967 data) input-output submodels for each of 23 sectors.

(3) Energy sources and conversion is represented by a linear process model for each of the 20 energy activities.

(4) In each period, a constraint equating supply and demand is imposed for every energy form.

(5) In each period, the requirement that supply equals demand for capital is introduced as the constraint on economic production.

(6) The model is solved as a linear program where total consumption is maximized subject to the energy and capital constraints and to initial and terminal capital constraints. The choice variables are the consumption-investment split and the level of each energy activity in each period.

The strengths of the model lie in the information that it yields about an efficient configuration of the energy supply and conversion sectors. The solution gives an efficient time path of the level of use of each energy activity including efficient time paths of fuel mix and energy technology choice. The demand for each fuel and technology takes account of the level and composition of economic activity, information on resource cost and availability, and the time-phasing of energy activity. Overall, the model provides a useful framework for the analysis of energy supply policy particularly those issues concerning the choice of technologies and fuels.

At the same time, the Pilot model has weaknesses and limitations. The model assumes that people optimize over a 40 year period, that they have perfect knowledge and foresight, and that behavior within each 5 year period is uniform; these are not realistic specifications from a behavioral or projection point of view. On the energy and economic side the model is very rigid – the economic structure is fixed, there is no role for prices, there is no interfuel substitution and no energy conservation. Thus, the representation of the economy and of energy is rather restrictive and serves essentially just to indicate the order of magnitude of energy demand.

From this survey, the useful range of applicability of the Pilot model can be indicated. Pilot is not a descriptive, nor a predictive, nor an economic model so it is not appropriate for economic projections or analysis. However, its economic content is sufficient to make the implied energy demands reasonable in terms of general magnitude. What the model does do well is the analyses of an efficient organization and evolution of energy supply to meet these energy demands. It is useful, therefore, for strategic analyses of broad energy supply and technology options.

IV. DRI Model

Data Resources, Inc. has constructed, and uses for short-run economic forecasting, a detailed macroeconometric model. This model gives comprehensive and detailed estimates of expenditure, price and financial variable and is in wide use for forecasting and economic analysis.

The structure of the model is essentially Keynesian, i.e. it is based upon estimates of demands or expenditures. Total expenditure yields GNP which is then disaggregated into income, industry and regional variables. Some of the features of the model are:

(1) Price and cost information is inserted into a "stage of processing model" to follow primary prices through to output prices.

(2) Monetary information is inserted into a financial model to obtain detailed estimates of financial conditions.

(3) Government activity, particularly taxes, purchases and transfers, is estimated in some detail.

(4) Estimates of prices, financial conditions, government activity and income are introduced into econometric expenditure functions to estimate consumption and investment purchases over many types of goods and services.

(5) Final demand spending – consumption, investment, government purchases and net exports – is summed to obtain GNP.

(6) GNP, equal to gross national income, is then disaggregated over income components. Also, final demand coupled with a fixed coefficient input-output system, is used to estimate industry outputs. Finally, output and income variables are allocated across geographical regions.

One strength of the DRI model is that a very large range and volume of information relevant to short run forecasting – price and cost trends, monetary and fiscal conditions, past behavior and current trends of consumer and business behavior, and judgemental information – is organized into a consistent framework. From this framework, a comprehensive and detailed set of forecasts is produced. These forecasts cover the macroeconomic aggregates as well as their implications for spending, industry, income and regional variables. The forecasts, in both macroeconomic and detailed form, are useful for business and government projections and planning.

At the same time, the model does have restrictions and limitations. As it is a Keynesian or expenditure-based model, it has no supply side and does not allow for productivity, capital and labor growth or supply constraints. While these may not be of primary significance in the short run, they can be of great importance in influencing the growth and structure of the economy in the medium and long runs. Thus, the model is a short run model. Also, the linkages from the macroeconomic side down to the detailed industry, regional and other variables are rigid and do not include any flexibility or behavioral content. While these linkages are appropriate as a disaggregating device, they are not appropriate for considerations involving structural change.

These features imply that the DRI model is very suitable for some applications but not for others. The model is not well suited to medium or long run projection, it is not a growth model, it is not well suited to the analysis of economic structure, nor is it an energy model. What it does do well is to provide a framework for comprehensive short run economic forecasts. It provides a meaningful basis for these forecasts and also expresses them, in consistent form, across a wide range of expenditure, industry, income and regional variables.

V. Hudson-Jorgenson Model

This is a model of the growth and structure of the U.S. economy. In its sectoral specification it emphasizes energy but it also covers the non-energy sectors of the economy. Spending and supply are modeled, as well as prices and quantities, to give an integrated representation of economic growth. The model has been used for projections and analyses of energy and economic growth and for the analysis of energy-economy interactions.

The structure of the model incorporates demand and supply aspects within an endogenous economic structure for a 10 sector model of U.S. economic growth. The components of the model include:

(1)   Prices of primary inputs, including capital and labor, are set so that demand and supply for these inputs are in balance. (This is achieved by iterating through the entire solution sequence.)

(2)   Output prices are calculated from the primary input prices, allowing for prevailing patterns of input into production and for production efficiency levels.

(3)   Product prices and factor incomes are introduced into final demand sub-models and these, together with information on government purchases and exports, generate estimates of the components of final demand expenditure on each type of good or service. The household behavior sub-model also calculates the labor supply (the labor-leisure split) and household saving (the spending-saving split).

(4)   Production sub-models, one for each sector, take the primary and intermediate input prices and sector efficiencies, to estimate input patterns, or input-output coefficients. These endogenous coefficients indicate the least cost feasible pattern of production, given prevailing prices and based on observed industry behavior.

(5) From real final demands and the input-output coefficients, industry outputs and interindustry transactions are calculated.

(6) This solution is repeated for every year with the economy growing through net investment, labor force increase and productivity increase.

The strengths of the Hudson-Jorgenson model are that economic interdependence and structural change is handled in a meaningful and consistent way and that supply, demand and other determinants of economic growth are consistently incorporated. The inclusion of the principal mechanisms of economic growth, on both demand and supply sides, makes the model suitable for medium and long run analyses of economic growth. This capability for economic projection and analyses is enhanced by the modeling of the structure of prices, spending and production. The explicit modeling of economic structure and growth provides a sound framework for the analysis of economic interdependence, particularly for analysis of energy demand and of energy-economy interaction. For example, the energy-economy interactions in the model allow for the effects of energy changes on general prices, input patterns and energy demands, on energy sector claims on labor, capital and other inputs, and on productivity GNP and economic growth.

At the same time, the model has limitations and areas of non-applicability. The behavioral and economic specification does not include short-run responses and adjustments so the model is not well suited to short-run analysis. The sectors are fairly broad so there is not a lot of detail in the spending and production information. Thus, the model is suited for aggregative but not for detailed analysis. The energy sectors have some detail in terms of output but do not have technological detail on the input side. This means that the model is not appropriate for energy supply analyses involving technological or process detail (although extension of the model by linkage to the Brookhaven TESOM model has bypassed this problem).

Given these features, the Hudson-Jorgenson model provides a good framework for some types of applications but is inappropriate for others. The model is not designed for, nor is it will suited to, short run pre-casting; it does not have extensive economic or energy detail; and it is not a framework for technology analysis. However, it does provide a sound framework for medium and long-run energy and economic projections, for energy-economy impact analysis, and for strategic evaluation of broad energy policies.

VI. Model Capabilities

This brief survey of the four models shows that each has different strengths and different areas of applicability. Although the models all relate to energy and/or the economy, each has principal relevance to only one part of this broad system. For summary purposes, an illustrative list of informational uses and applications of energy-economy models can be

drawn up. These applications, under the three categories of economic projections and analyses, energy projections and analyses, and energy-economy analyses, are set out in Table 1. This Table also sets out the four models, defining a matrix illustrating model capabilities.

The central feature of this matrix is that most of it is empty – no model caters to all the listed applications, each model has capability in only a few of the areas. This feature is also true of models other than the four surveyed here – any model at all has some applications it is suited to but many others for which it is inappropriate. In view of this, the conclusion follows that there is no all purpose model for energy or economic analysis.

This implies that, for any analysis, the model should not precede the problem – it is correct to have a model and to apply it to any and every problem that comes along. In other words, it is not correct to have an existing model and to force every problem through this model. This can be readily demonstrated from the Model-Capability Matrix by starting with any model, i.e. any column in the matrix, and noting that since many entries in that column are empty, there is no guarantee that the model will, if applied to a particular problem give valid or meaningful information. If a model is applied to the analysis of an inappropriate issue, then the resulting information is likely to be invalid and decisions based upon this information bad.

An example of the valid and invalid application of models can be given from recent work that we performed for the Department of Energy. The task was to compare the economic effects of a specified energy change estimated by the DRI model and by the Hudson-Jorgenson model. The same energy change was introduced into each model and the impact on GNP through 2000 was computed. For this application, involving energy-economy inter-action with long run adjustments in economic structure and growth, the Hudson-Jorgenson model is appropriate but the DRI model is not. The economic impacts estimated by the Hudson-Jorgenson model were consistently different from those estimated by DRI, in fact they were approximately twice the magnitude as those estimated by the DRI model. If the DRI model were the resident model and was applied to this long run impact problem, an inappropriate application, the resulting information would be wrong by a factor of two and any decisions based on this information would be sub-optimal. The use of this information would, by underestimating the costs of energy policy lead to the introduction of overly severe policy measures. (This comparison is reported in A Comparative Assessment of Energy-Economy Inter-actions, by R.J. Goettle, E.A. Hudson and J. Lukachinski, BNL 50923, Brook-haven National Laboratory, December 1978).

The correct strategy for analysis is not to proceed from the model to the problem but rather the reverse: the problem should dictate model selection. Given the differing model capabilities it is only valid to select a model once the objective, or the nature of the informational product required, is known. This means that the type of information re-quired has first to be defined, then the available models that can provide

Table 1

Model - Capability Matrix

Application	Almon	Pilot	DRI	Hudson-Jorgenson
1. Economic projections, analyses				
Level of activity, short run			X	
Detailed projections	X		X	
Economic growth				X
Economic structure				X
2. Energy projections, analyses				
Demand		X		X
Supply and conversion		X		
Conservation				X
Technologies		X		
3. Energy-econcmy analyses				
Short run impacts			X	
Full impacts				X

this information can be identified and the most appropriate model
selected. Only in this way can the probability be maximized that mean-
ingful information will be provided by the model framework. Correspond-
ingly, the expected value of decisions based upon model information will
be improved.

This strategy for model use can also be represented as a horizontal
approach within the Model-Capability Matrix of Table 1. The first step
in any analysis is to define the problem and the type of information
required to handle the problem. This information will correspond to one
row of the matrix. Consider, for example, the case in which information
is needed on energy demand. This row of the matrix shows that, of the
four models, only two offer relevant capability - the Pilot and the Hudson-
Jorgenson models. From these two, the most suitable can be selected, e.g.
if the demand information is needed in connection with supply decisions
then Pilot might be used whereas if the issue involved energy conservation
or price response, Hudson-Jorgenson would be more appropriate.

The information on model application might conveniently be stored
and presented in a problem-model mapping such as that given in Table 2.
Table 2 is an illustrative mapping, using the information on model
capabilities developed in the above survey, to provide a reference as
to which models might be appropriate to handle a specified problem. It is
a quick classification system that permits models appropriate for each
problem to be identified.

VII. Model Assessment and Selection

This review of illustrative energy-economy models and their
capabilities yields a series of guidelines for the assessment and appli-
cation of models. For assessment, it can be recognized, that since there
is no all-purpose model, it is neither valid nor useful to appraise any
model against the ideal of an all-purpose system. Rather, the primary
area of application of the model should be determined and the model
appraised according to first, whether it generates meaningful and useful
information within this area of applicability and second, whether infor-
mation within this area is useful and worth having. As part of the model
assessment it would also be useful to indicate the valid areas of applica-
tion and to record these in a form such as the Problem-Model Mapping.
This would give a ready indication, for potential users, as to whether
or not the model might be meaningfully applied to the areas or problems
in which they are interested.

For information and model agencies and for model builders, a similar
set of conclusions follows. It is not possible to build an all-purpose
model so resources should not be directed towards the construction of a
supermodel. Rather, the modeling strategy should be directed towards the
construction of a series of models, encompassing the areas of potential
interest for policy and planning. (These models can be constructed with

Table 2

Problem - Model Mapping

Problem	Model
Disaggregation to find specific economic impacts	Almon
Strategic choice in energy sources and technologies	Pilot
Short-run economic forecasts	DRI
Medium and long run economic analysis	Hudson-Jorgenson
Short run economic impacts of energy	DRI
Full economic impacts of energy	Hudson-Jorgenson

compatible interfaces to permit an extended overall model to be set
up in a modular fashion but even this can cover only a limited range
of the energy-economy system). Also, it would be useful for the modeling
agency to maintain a Problem-Model Mapping for its models. This would
provide a ready reference, from an applications point of view, of the
capabilities available and a guide as to which models would be useful
for any given application.

Finally, there is a set of conclusions for policy analysis. It is
important to always work from the problem to the model (i.e. to select
the model most suited to the problem at hand) rather than to try to use
only one or two models for all applications. The strategy for model selec-
tion passes through the steps:

· define the issue being addressed,

· specify the information product required to analyze the issue,

· select, from available models or analytical bases, the framework
  most suited to providing the required information product (a
  reference system such as the Problem-Model Mapping is a useful
  way of storing this information on model capabilities),

· apply the model to generate the range of information required.

                              DISCUSSION

Dr. Parikh: I would like to do some ventilation. I think essen-
tially what I would like to do is I want to make a few remarks and to
give you some additional reports that would very obviously change
your perceptions on the structure and content of the models developed
on the Pilot project.

Essentially, it seems to me that the particular model that you are
referring to is the model that was documented about two years ago in
a report published in October 1976 and, subsequently, the model was
used in the energy modeling forum exercise on energy and the economy.
Since then, a lot of work has gone on and, in particular, I would
like to draw your attention to some of the work I have done on some-
thing that I call the welfare equilibrium model and some of the key
features of this model are that it has a variable coefficient input/
output matrix. It is not a fixed coefficient structure. It is not
driven by a consumption objective, but it has something that I call
the household welfare function that explicitly takes into account
something that is usually called the labor/leisure trade-off. It also
includes resource supply curves mathematically, in a manner similar
to what is done in the SRI approach. I think these are very signifi-
cant differences in the model that you might be thinking of and the
current state of development on the Pilot project.

The thing that I would like to add is that it is an open question of whether the optimizing models and the simulation models really lead to very different answers, especially if the simulation models include some expectational variables. I think this is an area that requires some research and work. There is information embedded in optimizing models that, for all practical purposes, sort of comes out of the simulation models as well. After all, the simulation models are trying to optimize myopically and try to find some market equilibrium solutions.

Last, I think one of the other models you ought to include in your taxonomy here is Alan Manne's Eta-macro. I think that would enhance the list that you have because it is also an energy economy model that has been used quite a bit.

Dr. Hudson: Your points are well taken. I was using the latest model which I had documentation for. If we can spend time together, I would appreciate any more recent documentation.

With all due respect, I think the rest of your points are not relevant to the topic in the conversation which is, no matter what the specifics of the model are, it is going to have some areas of applicability and other areas of non-applicability. I accept that your model has been changed. For the same reason, I could have included Alan Manne's model--I haven't got much time. The point is simply, each model does something well and can't do other things.

Dr. Hogan (Harvard University): The suspense is killing me on the comparison between the DRI and the H-J model. We are clear as to which one was wrong but we didn't get the direction of the errors and, second, I am interested if you try to explain why you got different answers and if you could share that explanation with us?

Dr. Hudson: Well, let me say, I am only calling the DRI model wrong in this particular application. The direction of the error was that the long run GNP impact that we estimated were larger than the DRI one and, as for the reasons, there is a report due out of Brookhaven, either next week or the week after, which goes into them in some detail and I will be glad to give you a copy.

References

C. Almon, M.B. Buckler, L.M. Horwitz, T.C. Reimbold; <u>1985: Interindustry of the American Economy</u>; D.C. Heath; 1974.

M.L. Baughman, E.G. Cazalet, E.A. Hudson, D.W. Jorgenson, D.T. Kresge, E. Kuh, D.W. North; <u>Initiation of Integration</u>; Electric Power Research Institute, Palo Alto; 1978.

Data Resources, Inc.; <u>DRI U.S. Model Version 1978C</u>; DRI, Lexington, Mass.; 1978.

R.J. Goettle, E.A. Hudson, J. Lukachinski; <u>A Comparative Assessment of Energy-Economy Interactions</u>; BNL 50923, Brookhaven National Laboratory; 1978.

E.A. Hudson and D.W. Jorgenson; <u>The Long Term Interindustry Transactions Model</u>; Federal Preparedness Agency, Washington, D.C.; 1977.

S.C. Parikh; <u>Analyzing U.S. Energy Options Using the Pilot Energy Model</u>; Technical Report SOL76-27, Stanford University; 1976.

Econometric Models and Their
Assessment for Policy:
Some New Diagnostics Applied to Translog Energy Demand in Manufacturing

Edwin Kuh and Roy E. Welsch
Massachusetts Institute of Technology

Introduction

This paper has three parts. The first offers some perspective on the evolution of modeling during the past three decades and how this in turn has motivated greater interest in model evaluation. To attempt such a task in ten minutes or less verges on the presumptuous, yet I believe it may help to serve the purposes of this conference.

The second part introduces some new procedures for assessing the reliability of regression estimates by presenting some new diagnostic procedures that reveal which components of data have exceptional influence on estimated coefficients. The third part is an analysis of 3SLS estimates and extends the previous discussion to consider some limitations on using translog estimates in energy policy analysis.

I. A Bird's Eye View of Modeling

A. A Condensed Impression of Policy Modeling

When in the mid- to late 1940's econometric modeling as we now know it had its genesis at the Cowles Foundation, it was a time of exuberant optimism. I believe it is fair to say that the progenitors of modern economics believed that economic theory existed - or at the very least that the foundations were there to enable its successful evolution - to model reality well enough to encompass most if not all significant economic phenomena. Coupled with this strong faith in the power of economic theory was an equally optimistic belief in the power and adaptability of classical statistics to solve influential problems posed by economic theory.

Thirty years later these early high hopes for econometrics are muted. Econometric models often fall short of expectations, despite major gains in knowledge and many notable theoretical and empirical accomplishments. Some of these concerns, which will be spelled out more below, are a direct consequence of growing aspirations rather than just unfounded optimism.

Once engaged in the modeling process, it becomes evident that economic theory is silent or incomplete on some central issues, not the least of which are the provision of explicit guidance about dynamic structure, the formation of expectations (which in turn is related to dynamics), behavior of firms confronted by imperfect competition and what should be done in the presence of institutional change. Even where economic theory is sufficiently explicit, it is often difficult to make meaningful distinctions among alternatives through statistical inference. For example, production or consumption functions with rather different behavioral implications as a rule more often than not explain the data equally well. It is also clear that statistical theory, whether classical or Bayesian, does not correspond closely to the processes of iterative model building we all use.

445

Some of the shortcomings alluded to have their origins in the nature of data with which we must deal. In most instances the highly complex socio-economic processes whose essential nature we try to characterize are not amenable to controlled experimentation. Thus basic assumptions about exogeneity are not subject to verification and the isolation and measurement of theoretically significant components, in the absence of fortuitous experiments generated by Nature, are often difficult or impossible to capture by regression analysis in its various forms. Thus statistical paradigms which at least implicitly assume experimental control are often inadequate to their task.

If the preceding remarks are not too wide of the mark, it behooves modelers (of all methodological inclinations, not just econometricians) to make suitably modest claims for their progeny. We conduct our trade in circumstances where the most we can realistically hope to accomplish is an adequate approximation of reality. Even then, the quality of that approximation will diminish with time, often quite rapidly, while certain components will completely break down at times.

While experienced modelers know many limitations of their models, a verbal and written tradition of salesmanship has grown up which outsiders, including model clients, find hard to appraise. If we are smart and lucky at the same time, we can squeeze out some useful insights from models. At the same time, it is important to be explicit about the limitations of model analysis.

In an environment such as this, model assessment takes on increased importance. The development of procedures that measure the extent of model adequacy is badly needed if we are to make sensible claims about the information models can provide. Properly qualified, I am convinced that the benefits are both real and important.

B. The Changing Use of Models and Their Evaluation

Early on during the period surveyed, models were academic exercises, properly so since this art form was then in its infancy. Assessment consisted of peer review of the usual sort including exchanges in journals, conference proceedings and the like. During this time models were treated as esoteric mysteries by business and government alike. Beginning in the 1960's, it is our impression that models have come to be used, and sometimes even taken quite seriously, in the policy processes of government. The business sector currently pays handsome sums for short and medium term forecasts from which we might infer that they also take models seriously. The use of models in policy is qualitatively different from the use of models to advance knowledge for its own sake. In particular, the notions of what constitutes acceptable performance have changed (slowly, as the culture changes only gradually), and become tougher in certain respects.

While the application of models to policy problems was rapidly expanding, model structures were becoming more sophisticated and complex. Abetted by dramatic reductions in computer costs, modelers have attempted to mimic reality more closely, address more detailed and intricate policy issues and reduce aggregation biases. One major consequence of greater complexity has been to obscure the link between the user's intuition and model behavior. A black

box in the absence of powerful theory or strong intuitive comprehension does not inspire confidence. The commercialization of many models offers an institutional barrier with similar impediments to accessibility that permits peer review to function.

The overwhelming emphasis has been on the building and use of models, to the relative neglect of model evaluation. Users want answers yesterday for urgent matters and the model builders respond. Until model users insist on more than minimal evaluation, criteria and procedures for evaluation will continue to lag far behind.

For a host of reasons, including those already suggested, there has been growing scepticism and at times outright hostility to models. Given the nature of the social sciences, we should expect that models will continue to be inherently controversial. Appropriate model evaluation can lead to the articulation of explicit scientific standards for models and thus blunt some valid concerns about models in the policy process. When strengths and limitations have been realistically evaluated, models can offer useful insights into some policy issues more effectively than is now possible.

The remainder of this paper will discuss some statistical procedures that, among many others, can be useful in model evaluation, and that are related to some of the issues raised above. We would like, in the brief time at our disposal, to suggest an approach to applied econometric analysis that might alleviate some concerns that exist about the reliability of estimated equations in econometrics.*

II. Regression Diagnostics and Influential Data

The proposed methods are experimental in spirit. The basic process is to perturb slightly model inputs - which are construed broadly to include data and model error assumptions - by differences or derivatives, and to examine the influence of these perturbations on model outputs such as estimated regression coefficients, predicted or fitted values and standard errors. These procedures are designed to locate subsets of the data which exert an unusually large influence on model outputs. By doing so, two characteristics of much current econometric modeling can be improved. First, because model building is invariably an iterative process during which the initial model or data are frequently being modified, the standard test statistics reported at the end of the modeling process do not conform to the postulates of hypothesis testing. As we intend to demonstrate, the perturbation approach can reveal both strengths and weaknesses that elude classical test statistics in the ways that they are conventionally used. Second, since most econometric

data are non-experimental, standard techniques provide little, if any, insight into the problems for estimation that can thereby arise. The perturbation process does provide different and intuitively appealing information that is analogous to some aspects of experimentation.

Our general approach emphasizes both residuals and the X matrix together with its structure. To the extent that we are trained to view the X matrix as in principle "given", this might be considered heretical. Since in practice we are well advised to examine the X data for unusual components, the heresy is a minor one. We have systematized this process in the attempt to look for multivariate influential observations with especially strong effects on the model outputs.

Our goal, stated earlier, is to identify subsets of the data that appear to have a disproportionate influence on the estimated model and to ascertain which parts of the estimated model are most affected by these subsets. The relevance of this objective to the assessment of model adequacy is self-evident.

The sources of influential subsets are diverse. First, there is the inevitable occurrence of improperly recorded data. Second, observational or sampling errors are often inherent in the data and the diagnostics we propose below may reveal their unsuspected existence or severity. Third, outlying data points may be legitimate extremes. Such data often contain valuable information that improves estimation precision. Even in this beneficial situation, however, it is instructive to isolate extreme points and to determine to what extent the parameter estimates depend on them. Fourth, since the data could have been generated by model(s) other than that specified, diagnostics may reveal patterns suggestive of these alternatives.

Before describing multivariate diagnostics, a brief two-dimensional graphic preview in Exhibit 1 will indicate what sort of interesting situations might be detected.* Exhibit 1a portrays the ideal null case of a uniformly distributed dependent variable and, to avoid statistical connotations, what might be called an evenly distributed independent variable. The point o is anomalous in Exhibit 1b, but no adverse leverage effects are inflicted on the slope estimate since it occurs near the mean of the independent variable. The estimated intercept, however, will be affected.

Exhibit 1c illustrates an instance of leverage in which a gap arises between the main body of data and the outlier. Since this outlier is consistent with the slope information contained in the rest of the data, this situation may exemplify the benevolent third source of influence mentioned above in which the outlier supplies crucially useful information -- in this case, causing a reduction in variance.

Exhibit 1d is a more troublesome configuration that can arise in practice. In this situation, the estimated regression slope is almost wholly determined by the extreme point. Unless the extreme point is a crucial and valid piece of evidence, the researcher is likely to be highly suspicious of the estimate. Given the gap and configuration of the main body of data, the estimate surely has less than the usual degrees of freedom; in fact, it might appear that there are effectively only two data points.

---

* All Exhibits mentioned in this paper are collected at the end of this paper.

The leverage displayed in Exhibit 1e is a potential source of concern since o and/or ● will heavily influence the slope estimate, but differently from the remaining data. Here is a case where some corrective action is clearly indicated - either data deletion or, less drastically, a downweighting of the outliers or possibly reconsideration of the model.

Finally, Exhibit 1f presents an interesting case in which neither o by itself can affect the outcome very greatly. The potential effect of one outlying observation is clearly being masked by the presence of the other. This example serves as simple evidence of the need to examine the effects of more general subsets of the data.

Exhibit 2 indicates the standard notation which we use for ordinary least squares with the additional need to designate b(i) as the OLS estimate of β with the ith row of X and y data deleted. Exhibit 3 contains the usual OLS estimator and the projection matrix H whose diagonals h_i play a strategic role in subsequent analysis. In addition to noting the properties described in point 1, it is worth emphasizing point 2, namely that (for centered data) the h_i can be viewed as the multivariate distance of the vector x_i from its center. Thus large values of h_i serve as an indication of noteworthy leverage, an interpretation which is also consistent with viewing H as the projection of y into $\hat{y}$. For the bivariate case illustrated in point 4, it is transparently obvious that a large squared deviation $\bar{x}_i^2$ results in large h_i and large influence. Exhibit 3A contains analogous information about 2SLS.

Exhibit 4 presents several of the single row deletion formulae which are central to the diagnostics which will be discussed today and later illustrated with a translog production function for energy demand in manufacturing. Point 1 gives the formula for the difference between the full data set OLS estimates of β and those with the ith row deleted: This difference will be larger when both residuals and hat matrix diagonals (reflecting leverage) are larger. In the analysis of structural relations this is a measure of primary interest. To recognize inherent statistical variability, we have chosen to scale by the standard error of estimated β's using s(i) instead of s, which makes the numerator statistically independent of the denominator, along with the usual diagonal element of $(X^TX)^{-1}$. Thus each coefficient's sensitivity to a particular row of data can be determined and this can in turn be related to residuals ("studentized" as shown in point 5 as one desirable way to take account of scale), and leverage reflected by the hat matrix diagonals.

A second measure, for which unscaled and scaled versions appear in points 3 and 4 respectively, shows the influence of row deletion on the predicted value of y, designated by $\hat{y}$, when a row is deleted. It represents the combined effect of all coefficients and depends in a qualitatively similar way on e_i and h_i which also determine coefficient changes. The third and last measure, designated as COVRATIO in point 6, is the ratio of determinants of estimated parameter covariances for row deleted to complete-data-based estimation, which can be viewed as the ratio of two generalized variances. A small magnitude (taken to be less than 1-3 p/n), indicates that deleting the row improves overall precision of estimation as measured by the generalized variance, and large magnitudes (greater than 1+3 p/n) indicate that deleting

the row improves overall precision of estimation as measured by the generalized variance. Table 4A shows the comparable set of diagnostics for 2SLS that will be used in the following translog analysis.

All measures depend on hat matrix diagonals and residuals. How these can help to obtain a clearer understanding will be illustrated next. Considerable other experience suggests that additional and worthwhile insights about problem structure often emerge from applying these methods.

III. An Illustration: Translog Energy Demand in Manufacturing

The translog production function has come into vogue in recent years, particularly in the study of factor substitution in energy production where it forms the core of a growth model of Hudson and Jorgenson (1974). A four input version of the translog for manufacturing including capital, labor, energy and materials, called the KLEM model by Berndt and Wood (1975), is based according to standard assumptions of the Jorgenson model on symmetry, constant returns to scale, linear homogeneity in prices, perfectly competitive factor markets and long-run market equilibrium. The three resulting equations (the fourth is redundant since shares on the left-hand side add to one and its inclusion causes the error covariance matrix to be singular) appear below. Cost shares are the dependent variables which are functions of the same set of relative prices. The model is:

$$MK = AK+GKK*LOG(PK/PM)+GKL*LOG(PL/PM)+GKE*LOG(PE/PM) + \varepsilon_1$$

$$ML = AL+GKL*LOG(PK/PM)+GLL*LOG(PL/PM)+GLE*L)G(PE/PM) + \varepsilon_2$$

$$ME = AE+GKE*LOG(PK/PM)+GLE*LOG(PL/PM)+GEE*LOG(PE/PM) + \varepsilon_3$$

where MK, ML, ME are cost shares of capital, labor and energy respectively and PK/PM, PL/PM and PE/PM are capital, labor and energy price indexes relative to a materials price index.

Since computationally efficient diagnostics have not yet been devised for 3SLS (or other full information estimators) but do exist for 2SLS, we shall rely on the latter to provide initial diagnostic information about implications of model sensitivity and then rely on 3SLS estimates for long-run energy policy implications. In an attempt to answer these questions, at least partially, the following procedures were followed. First, all three equations (Capital Share, Labor Share and Energy Share) have been subjected to row deletion diagnostics. By way of example, more detailed discussion of 2SLS estimates for the Energy Share equation are presented. A subset of data which proved to be influential in the 2SLS version of all three equations was identified and the model was then re-estimated with iterative three-stage least squares, excluding the subset of influential data. While deletion and re-estimation is definitely not called for as a matter of course, it seems worthwhile in the context of model assessment to evaluate how particular components of the data affect reported measures of behavior, so we are following that procedure here. The implications of the removal of this subset of data for energy demand were evaluated by means of historical simulations and long-run extrapolations using three alternative assumptions concerning energy price behavior (no change, 50% increase and 100% increase).

A. Some 2SLS Details

The results of the Two-Stage Least Squares row deletion diagnostics are summarized in Exhibit 5.[*] Rather than show tables for each diagnostic-equation combination, a table identifies those points which were selected by means of cutoffs (discussed in Belsley, Kuh and Welsch (1979)) that flag 5% to 10% of the more influential data.[**] It is apparent by inspection that a relatively small subset of data is identified as influential for each equation and that substantial overlap exists in these subsets. The evidence points to 1947, 1948 and 1971 as observations which consistently exert disproportionate influence and hence warrant closer scrutiny.

A more detailed look at the energy share equation indicates how the deletion diagnostics can be used for individual equations. 2SLS estimates are of interest in their own right as well as one way of examining how much a priori restriction influence estimated coefficients.

Estimated results for just the energy share equation are shown in Exhibit 6. While the different estimation procedures ordinarily yield different estimates, coefficient magnitudes for 2SLS and 3SLS do not differ greatly, and the only corresponding terms with opposite signs have low statistical significance.

Exhibit 6 also shows the influence of row deletion on coefficient change corrected for scale, the DFBETAS. We use a cutoff of one-half to signal a noticeably large single row influence in the belief that if 1/25 of the data can cause a coefficient to change by 1/2 a standard deviation, it is worthy of attention. We observe that the initial year 1947 and the two terminal years have at least one coefficient which exceeds the criterion, while the remaining years do not. The substitution of capital for energy has been a matter of some interest, the elasticity of substitution depending critically on the estimate in this equation on the relative capital price term PK/PM. This coefficient seems especially sensitive in the last two years.

Exhibit 7 contains two essential elements for the interpretation of what we have observed above: tabulated values of h_i and scaled prediction residuals e_i^*. We observe that the terminal observation (1971) has both strong leverage and a sizable residual, 1970 has much more moderate leverage but a larger residual and finally, 1947 has leverage above the 2 p/n cutoff of .32 and a small scaled prediction residual.

We look more briefly at Exhibit 8 which shows two other diagnostics described earlier, DFFITS and COVRATIO. For DFFITS, the cutoff criterion of one-half contains similar information as that in the individual coefficient changes: 1947, 1970 and 1971 are years whose exclusion earlier raised a warning flag, while 1948 and 1968 are added. For COVRATIO, a large value (one exceeding 1+3p/n = 1.48) signals that the absence of this row increases the generalized variance of $s^2(X^TX)^{-1}$, so that its presence improves precision (and conversely for values less than the cutoff (1-3p/n = .52)), tells a somewhat different tale. Specifically 1948 has a beneficial effect on estimation precision, while 1958 has a harmful effect on estimation precision. From Exhibit 8, 1958 has one of the largest residuals while 1948 has a medium size residual and one of the larger hat matrix diagonals. It is not surprising to find that the recession year of 1958 has an adverse effect since the hypothesis of market equilibrium which underlies the translog relation is violated.

---

[*]We are greatly obliged to Berndt and Wood for making their data available to us.

[**]Some brief discussion of cutoffs for scaled predictive residuals, hat matrix diagonals and COVRATIO appears in Exhibits 3A and 4A respectively. The DFBETAS cutoff is described below.

B. Implications of Alternative 3SLS Estimates

The next step in the analysis involves re-estimation of the translog model with 3SLS using three different sample periods: the full Berndt-Wood sample (1947-1971), the full sample excluding 1947, 1948 and 1971 and finally, the full sample excluding only 1947 and 1948. The last sub-sample was chosen on the basis of the conjecture that *ceteris paribus*, older data are more suspect than are newer data. Iterative three-stage least squares was the estimation technique used to impose cross-equation constraints and to make the estimation results invariant to the equation dropped (see Berndt and Wood [1975]). The results of the estimation are reported in Exhibit 9. The most readily apparent change in the estimated coefficients is the decline of the cross-price terms, both in algebraic value and in relation to their respective estimated standard errors. Also quite noticeable is the rise in the energy own-price term. Both of these changes have important implications concerning the effect of a change in energy price on factor markets and consequently on the income shares of the four factors under consideration (capital, labor, energy and materials).

Historical simulations were used to obtain estimates of the impact of the different samples on the shares and on expenditures of the four factors in selected years, and on the partial elasticities of substitution between pairs of factors for 1970. These results are shown for selected years in Exhibit 10 and Exhibit 11. As can be seen from inspection of Exhibit 10, the choice of sample period made little difference to share predictions, seldom exceeding 2%. The impact on the elasticities of substitution may be obtained by comparing the different row entries in a particular column of Exhibit 11. Substantial sensitivity to the selection of sample period is exhibited by these elasticities of substitution. Their volatility is a direct reflection of coefficient sensitivity to the particular sample period.

While qualitative conclusions regarding complementary between energy and capital are unaffected in this instance, it would be little more than an act of faith to use estimated parameters from a relation as sensitive to individual elements of the data or estimators as point estimates for long-term energy policy analysis, as some (but not Berndt and Wood) have been wont to do.

The principal reason why predicted shares are so stable is apparent from looking at the estimated intercepts in Exhibit 9. The intercept terms in each of the equations have extraordinarily high magnitudes relative to their asymptotic standard errors. This in turn reflects the substantial stability in cost shares shown in the display of underlying data in Exhibit 12. Thus even when the coefficients for relative prices change substantially, their total contribution to the explanation of shares is so slight that predicted shares (and corresponding levels of expenditure) are only marginally affected. In a related vein, relative prices, with the exception of wages, did not vary greatly over the sample period which can also be seen in Exhibit 12.

To illustrate the relative insensitivity of predicted factor shares to the sum of the relative price effects, within sample simulations were performed and the deviation of predicted share values from the estimated intercept term was calculated for each of the three sample periods. This deviation is taken to be a relevant measure of the contribution made to the prediction of the factor shares by the sum of the relative price terms. Next

this difference is scaled by the predicted shares, the resulting figure being interpreted as the relative contribution of the sum of the relative price terms. Summary statistics of these magnitudes expressed as percentages are found in Exhibit 13. It is apparent that the contributions of the price terms are small. When one asks the question "small relative to what?" two responses emerge. First, it is our conjecture that the "absolute contributions" are small in the sense that the numbers are probably of the same order of magnitude as the error variance of the several price series. Second, all the percentage contributions have median values of 10% of less with a "grand median" of 6.5%. Thus the seeming paradox of stability of predicted shares across sample periods along with instability of price coefficients across sample periods is attributable to the fact that the regression intercept terms are responsible for most of the equations' predictive ability. If this conclusion is valid one must then conclude that, while the translog model can be used to predict within sample shares (which are easy to predict anyway), it is particularly ill-suited to policy analysis about relative price effects.

We turn next to some experiments on the sensitivity of shares and expenditures to the period of estimation when there are large variations in the relative price of energy. Relative input price variability is introduced by entering "large" changes in energy prices which for present purposes are assumed to be a 50% increase and a 100% increase in the relative price of energy. The impact of these shocks is investigated by making out-of-sample extrapolations of shares and expenditures to 1985 and 2000 for each of the two assumed price paths in combination with each of the three estimated sets of parameters.

Two alternative growth rates (3% and 6%) of total output, which seems a reasonable range (however uncomfortably wide) to consider realistically, are assumed in calculations for expenditure figures. The energy expenditure figures in Table 14 show more sensitivity than either the capital or the labor expenditure figures. The full-sample period estimates of energy consumption in manufacturing (row one of each triplet) are well below those of either of the other two sample periods, which lie close together. If we believed the full sample estimates and the growth in energy demand was 3%, our point estimates of energy expenditures will be $86.08 billion in the year 2000, with energy price doubled, while the other estimates are $13-$16 billion larger. The span of interval estimates would of course be much greater. Other evidence on the sensitivity of the choice of sample period for long-run energy projections can be found in Exhibit 11 which contains the estimated partial elasticities of substitution for 1970 values of relative prices.

By way of summary, we may ask: What are the implications of this evidence for long-run energy policy analysis? A certain minimal lesson to be learned from these exercises is that uncritical application of this and similar models in projecting the impact of relative price changes on energy demand is likely to give a misleading sense of certitude. Regression diagnostics of the sort proposed here can yield information about parameter values that help in understanding the limitations of quantitative forecasts. Additionally, one may doubt the wisdom of using a simple equilibrium model such as the translog to answer questions concerning the impact of radical exogenous changes in relative prices.

DISCUSSION

Dr. Nissen: Well, I want to go on record to be the first to say that these techniques of regression diagnostics seem to me to be extremely useful, in an interesting way, to analyze the data that underlies the statistical model and to create or cast a greater insight into just what determines the empirical structure of a model.

In connection with the KLEM Model, I think it's appropriate to point out that the techniques have just been developed for single equation in two stage, well, single equation methods. When you check these results by taking the clue offered by the diagnostic that 1971 is an especially interesting year, and then pursue actually dropping that year from the model, and using the simultaneous equation methods--that is, preserving such things, imposing on the model such things as symmetry, which is difficult to explain the lack of within any kind of reasonable economic theory. It turns out not to have made that much difference. So while the method is very suggestive in terms of things to pursue further, in this particular case, it doesn't raise as much doubt about our result, upon further analysis than Ed's comments might suggest.

Bibliography

Belsley, David A., Edwin Kuh and Roy E. Welsch (1979), Regression Diagnostics: Identifying Influential Data and Sources of Collinearity, (to be published by John F. Wiley & Sons).

Berndt, Ernst R. and David O. Wood (1975), "Technology, Prices, and the Derived Demand for Energy", Review of Economics and Statistics, Vol. LVII, No. 3, Nov. 1975, pp. 259-68.

Hoaglin, David C. and Roy E. Welsch (1978), The Hat Matrix in Regression and ANOVA", The American Statistician, February.

Hudson, Edward A. and Dale W. Jorgenson (1974), "U.S. Energy Policy and Economic Growth, 1975-2000", The Bell Journal of Economics and Management Science, Autumn, pp. 461-514.

Phillips, G.D.A. (1977), "Recursions for the Two-Stage Least-Squares Estimator", Journal of Econometrics, pp. 65-77.

Rao, C.R. (1965), Linear Statistical Inference and Its Applications, New York: Wiley.

EXHIBIT 1

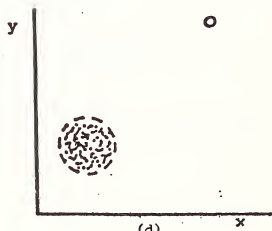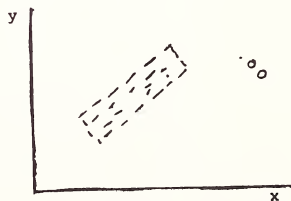Plots for Alternative Configurations of Data

(a)

(b)

(c)

(d)

(e)

(f)

457

EXHIBIT 2

Notation for OLS Estimation

Population Regression $y = X\beta + \varepsilon$	Estimated Regression $y = Xb + e$
y : nxl column vector for dependent variable	same
X : nxp matrix of explanatory variables	same
β : pxl column vector of regression coefficients	b : estimate of β
ε : nxl column error vector	e : residual vector

Addition notation

x_i : i^{th} row of X matrix	same
σ^2 : error variance	s^2 estimated error variance
	$b(i)$: β estimated with i^{th} row of X data matrix and y vector deleted.

Other notation is either obvious or will be introduced in its specific context.

EXHIBIT 3

Ordinary Least Squares and the Hat Matrix

$b = (X^TX)^{-1} X^Ty$

$\hat{y} = Xb = Hy$ where $H = X(X^TX)^{-1} X^T$

1. The projection matrix H has diagonals h_i with the following properties:

$$0 \leq h_i \leq 1$$

$$\sum_{i=1}^{n} h_i = p \qquad \text{(since X is of full rank)}$$

2. The projection matrix diagonals can be viewed for centered data as the distance of x_i from $\overline{x}$:

$$\tilde{h}_i \text{ (centered)} = h_i - \frac{1}{n} = \tilde{x}_i (\tilde{X}^T\tilde{X})^{-1} \tilde{x}_i^T \qquad (\tilde{x}_i \equiv x_i - \overline{x})$$

3. The average size of h_i is p/n. If X's are assumed to be joint normally distributed, it can be shown that for p>10 and n-p>50, $(n-p)(h_i - \frac{1}{n})/(1-h_i)(p-1)$ is $\sim F(p-1, n-p)$ with a value of about 2 at the 95% significance level. Hence we consider $h_i >, 2p/n$ to be a leverage or influential point i.e., $\tilde{x}$ is far from $\tilde{X}$ in a multivariate sense.

4. By way of illustration, for the bivariate regression case,

$$h_i = \frac{1}{n} + \frac{\tilde{x}_i^2}{\sum_{i=1}^{n} \tilde{x}_k^2} .$$

Reference: See Hoaglin and Welsch (1978).

Exhibit 3A

Two-Stage Least Squares and the Hat Matrix

The Model: $\qquad y_1 = Y_1\beta + X_1\gamma + \varepsilon_1 = Z\delta + \varepsilon_1.$ $\qquad (\varepsilon_1 \text{ iid})$

$$d = (Z^T X(X^T X)^{-1} X^T Z)^{-1} Z^T X(X^T X)^{-1} X^T y_1 = (Z^T H_x Z)^{-1} Z^T H_x y_1$$

$\qquad \hat{Z} = H_x Z$ and since H_x is idempotent we have

$$= (\hat{Z}^T \hat{Z})^{-1} \hat{Z}^T y_1$$

$$\hat{y}_1 = Zd = Z(Z^T H_x Z)^{-1} Z^T H_x y_1 = T y_1$$

(1) The projection matrix T, unlike the projection matrix H in the OLS case, is not symmetric nor are the diagonals of T bounded either from above or below. However, like H it is idempotent and its trace equals its rank.

(2) The diagonals of the projection matrix based $\hat{H} = \hat{Z}(\hat{Z}^T \hat{Z})^{-1} \hat{Z}^T$ are defined as h_i and those of the projection matrix T are defined as τ_i.

(3) An alternative projection matrix which is most relevant for second stage estimation should be used to assess the leverage effects of individual data points. The matrix $\hat{H}$ is symmetric and idempotent, therefore having trace = p, the number of regressors, and consequently having p/n as the average value of its diagonal elements. The use of this matrix derives from the "repeated least squares" interpretation of two-stage least squares and from the use of the h_i as measures of influence in the OLS case. A value of $\hat{h}_i$ greater than 2p/n then implies that observation is a leverage point.

(4) The projection matrix diagonal can be viewed for the centered data as the distance of z_i from $\bar{z}$.

460

EXHIBIT 4

OLS Single Row Deletion Diagnostics

1. Change in slope coefficients (DFBETA)

$$b-b(i) = (X^TX)^{-1} x_i^T e_i/(1-h_i)$$

2. Scaled change in slope coefficients (DFBETAS)

$$\frac{b_j - b_j(i)}{s(i)\sqrt{(X^TX)^{-1}_{jj}}} \quad,$$

where the usual estimate of σ^2, s^2, is replaced by

$$s^2(i) = \frac{1}{n-p-1} \sum_{k \neq i} (y_k - x_k b(i))^2$$

which makes the denominator stochastically independent of the numerator

3. Change in $\hat{y}_i$ (DFFIT)

$$\hat{y}_i - \hat{y}_i(i) = x_i(b-b(i)) = \frac{h_i e_i}{1-h_i}$$

4. Scaled changed in $\hat{y}$ (DFFITS)

$$\frac{x_i(b-b(i))}{s(i)\sqrt{h_i}} = \sqrt{\frac{h_i}{1-h_i}} \quad \left(\frac{e_i}{s(i)\sqrt{1-h_i}}\right)$$

Note: The factor $\sqrt{h_i}$ corrects for the fact that the fit does not have a scalar covariance matrix.

5. Scaled predicted residual or studentized residual

$$e_i^* = \frac{y_i - x_i b}{s_{(i)}\sqrt{1+x_i(X_{(i)}^T X_{(i)})^{-1}x_i^T}} = e_i/s_{(i)}/\sqrt{1-h_i}$$

Note: If ε is normally distributed e_i^* is distributed as t with n-p-1 degrees of freedom.

461

6. Ratio of single row deleted estimated coefficient covariance determinant to that for the full data set (COVRATIO)

$$\frac{\det s^2(i) \ (X_{(i)}^T X_{(i)})^{-1}}{\det s^2 \ (X^T X)^{-1}} = \frac{1}{(\frac{n-p-1}{n-p} + \frac{e_i^{*2}}{n-p}) \ (1-h_i)}$$

Note: This can be viewed as the ratio of two generalized variances and used as an overall measure of a particular row's influence on estimation precision. Magnitudes outside $1 \pm 3 \ p/n$ are indications of strong influence.

7. In all these various measures, the residuals e_i (often studentized) and hat matrix diagonals h_i appear as basic quantities. Either one alone does not suffice.

8. The fundamental deletion formula is known as the Sherman-Morrison-Woodbury Theorem (Rao, 1965, Problem 2.8, p. 29):

$$(X(i)X(i))^{-1} = (X^T X)^{-1} + \frac{(X^T X)^{-1} x_i^T x_i (X^T X)^{-1}}{1-h_i} \tag{1}$$

Exhibit 4A

2SLS Single Row Deletion Diagnostics

(1) Scaled change in slope coefficients (DFBETAS)

$$\left. d_j - d_j(i) \middle/ s(i) \sqrt{(\hat{Z}^T\hat{Z})^{-1}_{jj}} \right.$$

Note: This is based on full row deletion i.e., for all data in both
the first and second stages rather than just deleting $[\hat{X}_{1i} \mid X_{1i}]$
and ignoring the first stage.

(2) Change in $\hat{y}_i$ (DFFIT)

$$\hat{y}_i - \hat{y}_i(i) = z_i(d - d_{(i)})$$

(3) Scaled change in $\hat{y}_i$ (DFFITS)

$$\frac{z_i(d - d(i))}{s(i) \sqrt{z_i\, (\hat{Z}^T\hat{Z})^{-1}\, z_i^T}}$$

Note: The scaling term is an estimate of the standard deviation of $\hat{y}_i$
derived from the limiting distribution of $z_i(d - \delta)$

(4) Scaled predicted residual

$$e^*_i = \frac{y_i - z_i d(i)}{s(i) \sqrt{1 + z_i[Z_{(i)}^T X_{(i)} (X_{(i)}^T X_{(i)})^{-1} X_{(i)}^T Z_{(i)}]^{-1} z_i^T}}$$

Note: The scaling term s(i) is an estimate of the standard deviation of the
forecast error where the forecast in question is the two-stage
least squares forecast of y_i using all data but observation i in
the calculation of the parameters.

If ε is normally distributed then e^*_i is asymptotically normally
distributed.

463

(5) The ratio of the determinant of single row deleted estimated coefficient covariance matrix to that for the full data set. (COVRATIO)

$$\frac{\det\ (s^2(i)(\hat{Z}_{(i)}^T\hat{Z}_{(i)})^{-1})}{\det\ (s^2\ (\hat{Z}^T\hat{Z})^{-1})}$$

Note: As in the OLS case COVRATIO may be viewed as the ratio of the generalized variances and used as a measure of a particular row's influence on the precision of the estimation. Values of CORATIO outside of $1 \pm 3\ p/n$ are indications of strong influence.

(6) Two-stage least squares recursion formulae may be found in G.D.A. Phillips (1977).

EXHIBIT 5

2SLS Diagnostics: Influential Points Identified by External Scaling Criteria

Date	$\hat{h}_i$	Scaled Predicted Residual	DFFITS	COVRATIO	DFBETAS Intercept	β_{KK}	β_{KL}	β_{KE}	β_{LE}	β_{LL}	β_{EE}
1947	K,L,E		K	K,L	K,E		K		E		E
1948		L		K,E			L	K	L		
1949				L							
1950		L									
1951				L							
1958		K		E							
1970								K,E			
1971	K,L,E	K,L	L	L	K,L	K	K,L	K,L	E,L	L	

Note: K Represents Capital Share Equation

L Represents Labor Share Equation

E Represents Energy Share Equation

465

EXHIBIT 6

A. 2SLS Estimates of Translog Energy Demand Equation for Manufacturing

$$ME = .0414 + .0527PG \text{ LOG } PE/PM - .0074 \text{ LOG } PK/PM + .0019 \text{ LOG } PL/PM$$
$$\quad (23.4932)(3.444) \qquad\qquad (-.1440) \qquad\qquad (.449)$$

$$s = .0020 \quad R^2 = .6476 \quad DW = 1.6328$$

3SLS estimates (from Berndt and Wood)

$$ME = .0442 + .0214 \text{ LOG } PE/PM - .0102 \text{ LOG } PK/PM - .0043 \text{ LOG } PL/PM$$
$$\quad (38.078) \quad (2.343) \qquad\qquad (-2.444) \qquad\qquad (-1.438)$$

(t statistics appear beneath each estimated coefficient)

B. 2SLS DFBETAS: Translog Energy Demand

Year	Constant	GEE	GKE	GLE
1947	*-.7973	*.7807	-.0941	*.7368
1948	-.0867	-.3719	-.3538	.0604
1949	-.1158	.0510	-.2347	.0631
1950	-.0278	-.0410	.0480	.0418
1951	* -.4993	.4625	-.1735	.4393
1952	-.0760	.0447	.1871	.0994
1953	-.1480	.0483	.1407	.1427
1954	-.0624	.2121	.1329	.0877
1955	* -.4984	.4148	-.0428	.4484
1956	-.3619	.4111	.1653	.3542
1957	.0368	-.0771	-.2378	-.0475
1958	-.3828	* .5234	-.0057	.4101
1959	.0166	.0988	.1580	.0406
1960	.1928	-.0832	.1583	-.1140
1961	.0533	.0466	.1231	.0447
1962	.1455	-.1181	.0527	-.0787
1963	-.0052	.1183	.2257	.1136
1964	-.0001	-.0189	-.0339	-.0017
1965	-.1418	.1667	-.0003	.0772
1966	-.0957	-.0557	-.3573	-.0345
1967	-.0498	.0959	.0480	.0211
1968	.0602	-.1229	-.3126	-.2567
1969	-.1160	.1868	.0547	.0692
1970	-.3372	-.2183	*-.7450	.4326
1971	* .5625	.1012	*.8753	*-.6631

466

Exhibit 7

2SLS Hat Matrix Diagonals and Scaled Prediction Residual: Translog Energy Demand

Year	$\hat{h}_i$	τ_i	Scaled Prediction Residual
1947	*.3429	.5049	−.0979
1948	.3034	.5651	−.7646
1949	.1917	.2624	.2182
1950	.1633	.0840	−.5732
1951	.1551	.1810	.0995
1952	.1015	.0973	−1.0142
1953	.0668	.0626	−1.6408
1954	.1134	.0926	.6519
1955	.0883	.0301	.8998
1956	.2101	.1644	−.6195
1957	.0925	.0909	1.0069
1958	.1395	.0387	1.6142
1959	.0746	.0516	.6725
1960	.0466	.0327	.9086
1961	.0560	.0441	1.3928
1962	.1983	.0620	.8064
1963	.1019	.0919	.9929
1964	.2074	.1465	.0472
1965	.0943	.0911	−.6911
1966	.2156	.1602	−.7208
1967	.1244	.1117	−.3547
1968	.1174	.1095	−1.3101
1969	.1282	.1387	−.4358
1970	.2555	.2588	1.7874
1971	*.4102	.5261	−1.4716

Exhibit 8

2SLS DFFITS and COVRATIO: Translog Energy Demand

Year	DFFITS	COVRATIO
1947	*-.7973	.9886
1948	*-.5563	*2.6036
1949	.2731	1.4580
1950	-.1850	1.1103
1951	-.3591	.8807
1952	-.3301	1.0725
1953	-.4204	.7451
1954	.2586	.9929
1955	.0677	.6340
1956	-.2187	.8392
1957	.3256	1.1298
1958	.2859	*.4215
1959	.2169	1.0442
1960	.2575	1.2074
1961	.2975	.8430
1962	.2019	1.0634
1963	.3184	1.0055
1964	-.0126	1.3077
1965	-.1394	1.0003
1966	-.3654	1.2109
1967	-.1372	1.1929
1968	*-.5724	1.0014
1969	-.2398	1.1168
1970	*.9287	.8500
1971	*-.9723	1.2709

EXHIBIT 9

Iterative 3SLS Estimates of Complete Translog Model For Alternative
Sample Periods

Sample Periods

coefficient	1947-1971	1949-1970	1949-1971
Intercept(K)	.0563 (.0014)	.0599 (.0019)	.0577 (.0016)
γ_{KK}	.0248 (.0071)	.0387 (.0081)	.0312 (.0071)
γ_{KL}	.0003 (.0040)	-.0071 (.0049)	-.0021 (.0041)
γ_{KE}	-.0101 (.0040)	-.0067 (.0052)	-.0084 (.0045)
Intercept(L)	.2538 (.0028)	.2489 (.0019)	.2511 (.0019)
γ_{LL}	.0738 (.0067)	.0876 (.0056)	.0801 (.0057)
γ_{LE}	-.0042 (.0028)	-.0049 (.0052)	-.0008 (.0039)
Intercept(E)	.0441 (.0010)	.0447 (.0021)	.0427 (.0015)
γ_{EE}	.0213 (.0083)	.0287 (.0142)	.0377 (.0153)

Note: Coefficient standard errors in parentheses

EXHIBIT 10

Predicted Shares and Expenditures From Within Sample Simulations

A. Cost Shares

Equation:	date	Sample Periods		
		1947-1971	1949-1970	1949-1971
Energy:	1950	.047	.047	.046
	1959	.044	.044	.046
	1974	.045	.045	.047
Capital:	1950	.051	.051	.051
	1959	.056	.058	.057
	1971	.049	.046	.048
Labor:	1950	.260	.258	.259
	1959	.275	.273	.274
	1971	.296	.301	.298

B. Expenditures

Equation:	date	1947-1971	1949-1970	1949-1971
Energy:	1950	10.50	10.49	10.47
	1959	15.73	15.89	15.83
	1971	29.99	29.83	30.73
Capital:	1950	11.27	11.37	11.28
	1959	20.11	20.71	20.39
	1971	32.73	30.04	31.68
Labor:	1950	57.74	57.27	57.43
	1959	98.42	98.05	98.12
	1971	194.95	198.22	196.09

EXHIBIT 11

Energy Price Assumptions

Allen Partial Elasticities of Substitution: 1970 Values

Sample Period	Baseline	50%	100%
	σ_{KL}		
1947-1971	1.024	1.0266	1.028
1949-1970	.528	.499	.476
1949-1971	.859	.849	.841
	σ_{KE}		
1947-1971	-3.495	-3.038	-2.828
1949-1970	-2.171	-1.612	-1.352
1949-1971	-2.890	-2.034	-1.676
	σ_{LE}		
1947-1971	.652	.709	.740
1949-1970	.669	.740	.774
1949-1971	.929	.948	.956

EXHIBIT 12

Factor Shares, Relative Factor Prices and Their Respective Coefficients of Variation
(Expressed as fractions)

Observation	Energy Share	Capital Share	Labor Share	PE/PM	PK/PM	PL/PM
1947	0.0425	0.0510	0.2472	1.0000	1.0000	1.0000
1948	0.0512	0.0583	0.2771	1.2343	0.9535	1.0941
1949	0.0507	0.0461	0.2590	1.1265	0.7023	1.0881
1950	0.0460	0.0499	0.2479	1.0801	0.8241	1.0987
1951	0.0448	0.0505	0.2548	1.0286	0.8642	1.0993
1952	0.0445	0.0492	0.2665	1.0663	0.3333	1.1499
1953	0.0436	0.0474	0.2682	1.0710	0.8478	1.2050
1954	0.0478	0.0565	0.2716	1.0807	0.9047	1.2052
1955	0.0451	0.0527	0.2646	1.0843	0.8937	1.2203
1956	0.0457	0.0461	0.2687	1.0604	0.7723	1.2230
1957	0.0481	0.0504	0.2717	1.0559	0.8159	1.2596
1958	0.0483	0.0603	0.2727	1.0500	0.8740	1.2614
1959	0.0456	0.0620	0.2729	1.0457	1.0030	1.3261
1960	0.0458	0.0580	0.2773	1.0304	0.9391	1.3309
1961	0.0463	0.0591	0.2783	1.0246	0.9432	1.3548
1962	0.0452	0.0559	0.2827	1.0218	0.9416	1.3991
1963	0.0446	0.0561	0.2796	1.0119	0.9964	1.4524
1964	0.0439	0.0546	0.2833	1.0279	0.9852	1.4867
1965	0.0411	0.0548	0.2799	1.0079	1.0257	1.4943

EXHIBIT 12 (Con'd)

Observation	Energy Share	Capital Share	Labor Share	PE/PM	PK/PM	PL/PM
1966	0.0401	0.0547	0.2835	0.9874	1.0258	1.5043
1967	0.0407	0.0535	0.2867	0.9773	0.9566	1.5489
1968	0.0397	0.0576	0.2888	0.9993	1.0460	1.6299
1969	0.0396	0.0542	0.2902	0.9551	0.9476	1.6068
1970	0.0434	0.0536	0.2972	0.9428	0.8753	1.6834
1971	0.0447	0.0468	0.2890	1.0626	0.7770	1.7810
$\hat{\sigma}/\bar{X}$	0.0692	0.0837	0.0468	0.0565	0.0988	0.1574

EXHIBIT 13

Summary Statistics for Percentage Contribution of Relative
Price Terms to Predicted Factor Shares

Equation: Period/statistic	Median	Maximum	Minimum	Interquartile Range
Energy: 1947-1971	4.31%	6.77%	.12%	3.05% – 5.43%
1949-1970	4.4%	7.0%	.27%	2.45% – 5.37%
1949-1971	3.9%	9.47%	.36%	2.37% – 8.2%
Capital: 1947-1971	3.93%	14.03%	.11%	1.25% – 9.15%
1949-1970	10.0%	31.4%	3.0%	6.4% – 16.5%
1949-1971	6.5%	20.0%	.01%	2.7% – 11.9%
Labor: 1947-1971	7.85%	14.27%	2.5%	5.3% – 10.6%
1949-1970	9.5%	17.3%	3.5%	6.7% – 12.5%
1949-1971	8.6%	15.69%	3.0%	6.0% – 11.5%

Grand Median: 6.5%

EXHIBIT 14

Extrapolation of Energy Consumption in Manufacturing to 1985 and 2000:
Alternative Energy Price and Growth Assumptions
$ Constant 1970 Billions

ENERGY

	Baseline		50%		100%	
	1985	2000	1985	2000	1985	2000

Growth Rate = 3%

1947-71	40.88	63.70	49.29	76.79	55.25	86.08
1949-70	40.16	62.57	55.40	86.32	63.45	98.96
1949-71	40.06	62.42	54.94	85.60	65.50	102.05

Growth Rate = 6%

1947-71	62.89	150.73	75.82	181.82	84.99	203.70
1949-70	61.78	148.06	85.24	204.27	97.61	233.94
1949-71	61.63	147.70	84.52	202.55	100.76	241.47

CAPITAL

Growth Rate = 3%

1947-71	52.38	81.61	48.37	75.36	45.52	70.92
1949-70	50.05	77.97	47.39	73.83	45.50	70.89
1949-71	51.52	80.26	48.17	75.05	45.79	71.35

Growth Rate = 6%

1947-71	80.59	193.13	74.41	178.33	70.03	167.83
1949-70	76.98	184.49	72.89	174.70	69.99	167.74
1949-71	79.25	189.92	74.09	177.58	70.45	168.83

LABOR

Growth Rate = 3%

1947-71	284.49	443.21	282.80	440.59	281.60	438.72
1949-70	287.65	448.14	286.05	445.66	284.92	443.89
1949-71	285.12	444.21	284.68	443.68	284.54	443.30

Growth Rate = 6%

1947-71	437.62	1048.78	435.02	1042.56	433.18	1038.15
1949-70	442.49	1060.44	440.03	1054.56	438.20	1050.38
1949-71	438.60	1051.13	438.08	1049.87	437.71	1048.99

ON A PERSPECTIVE FOR ENERGY MODEL VALIDATION

Lawrence S. Mayer

Department of Statistics
Princeton University

1. INTRODUCTION

A recent article in the <u>Chronicle of Higher Education</u> suggests that one indicator of the health of a discipline is the proportion of its scholars regarded as "big thinkers." I am pleased to report that by this criterion, model validation is in marvelous health. The previous presentations have convinced me once again that we are blessed with a copious supply of splendorous notions of validation and its uses. But although this state of affairs may be a sign of healthy originality, it masks a number of fundamental problems, including a lack of formal, rigorous definitions of the basic terms we employ.

A group of scholars, no matter how sophisticated, is not likely to agree on relationships between ill-defined concepts, and, moreover, such agreement would be meaningless in the absence of basic definitions. We are often far too eager to define the relationships between grand concepts such as Validation, Evaluation, Verification, and Ventilation before we have even agreed on the meanings of these terms. I would be the first to admit -- and I am sure that many here will bear witness -- that I do not know the complete meaning of the simplest of these terms, "validation." My goal here is to introduce a perspective on validation that will make absolutely and explicitly clear its assumptions and definitions -- the precondition of reasoned debate and credibility in the scientific and political worlds in which we operate. I ask my colleagues, then, to be tolerant, rest their minds, and listen to a few ruminations on this question of perspective.

Traditionally, we have paid a great deal of solicitude to the question: "Are modelers doing a good job?" And after weighing the evidence, most of us have concluded that given the current state of information about energy processes, modelers are doing as well as can be expected. I endorse this conclusion but suggest the question itself is misleading because it confuses and confounds issues vital to the energy analyst struggling to assess the value of a model.

In particular, the question as framed ignores the dual nature of energy models as products of science and agents of policy. These natures are often confused; they must be assessed separately using distinct sets of criteria, outlines of which I will explore in this paper.

Pretend for a moment, if you will, that as modelers (and validators are all modelers in another incarnation) our purpose is not to convince our sponsors to increase our funds, not to persuade our academic brethren and sistren of the legitimacy of our discipline, not to secure fame and glory, and not to propagate our species by recruiting and training students -- but actually to estimate the scientific and political merits of our products. Stripped of extraneous issues, then, the question of validation divides in two:

A. To what extent, and in what way, does an energy
    model teach us about the world we live in?

B. To what extent, and in what way, is an energy model
    an important agent in the political process?

In answering question A, we find ourselves in the realm of the
energy model as science. We would not pose the question at all if we
did not believe, as scientists, that the stronger the scientific under
pinnings of a model, the more likely it is to tell us something about
the world. To answer the question, we must judge the model by the
degree to which its methodology conforms to the spirit and canons of
contemporary science. It is wrong, and probably fatal, to continue
to apologize for serious violations of scientific method. We can,
should, and must withstand the scrutiny of our peers in science.

Question B, on the other hand, takes the model from the womb
of scientific creation into the world of political life. To evaluate
a model as an actor in this world, we must isolate its uses as a polit-
ical tool, judge the aptness of these uses, and estimate the ability
of the model to contribute to political victories. To start with, we
must ask how the model's advertised accuracy compares with the actual
needs of those who commissioned it. But as a supplement, we must also
ask about its other roles, which, though not based in science, can be
recorded and assessed scientifically. For example, we can and should
assess the degree to which political debate is enhanced or restricted
by the use of a particular model. Similarly, the political implica-
tions of the perspective underlying a model should be appraised, since
all such perspectives contain political suppositions and biases. We
can no longer afford to dismiss critics of the political uses of model
as anti-technological cranks.

And so, as much as I welcome our discipline's day in the court
of science, I also encourage its appearance in the chamber of public
policy. Just as we must examine the scientific integrity of our work
by probing what it can actually tell us about the world, so we must
examine its political integrity by asking to what political uses it
is adaptable and to which misuses it is vulnerable.

All the while, we must bear in mind that the answer to one of
these questions does not necessarily follow from the answer to the
other. The best scientific model can prove unacceptable for a var-
iety of political reasons: it may not provide policy makers with
the forecasts they need, those they expected, or those which are
robust against unforeseen changes in exogenous variables. And con-
versely, the least scientific of soothsayers, acting to satisfy a
policy maker's needs, could produce a forecasting mechanism that
turns out to be accurate.

As scientists, we believe that on the average, science is the best approach for modeling empirical processes. The evidence for this belief lies in the aggregate, however. It does not and cannot rest on the supposition that science will produce an accurate forecasting model for each and every problem. It is this dual perspective on validation that I intend to develop in this paper and to supplement with an example drawn from the evaluation of a single-equation model for the rate of production of domestic crude oil.

The establishment of validation as a legitimate enterprise involves questions in need of serious study. Before anyone begins to repine, let me admit that I have no complete answers. But if the most critical step in understanding a process is focusing attention on the most pertinent questions, and if the process at hand is energy modeling, then these are the questions to work on. Asking them will not replace or displace the work of my colleagues, but it will require work different from that being reported here or, in fact, any currently funded by the Department of Energy. Answering these questions addresses issues that must be considered if energy modeling is to develop what it so sorely lacks: the cumulative nature of a true scientific enterprise and the accompanying respect of policy makers, policy analysts, and, most importantly, the public at large.

2. PERSPECTIVE

The perspective on model validation that I have tendered draws heavily on three experiences. The first is having been a Principal Investigator of the Department of Energy-sponsored Princeton Residential Energy Conservation Project, a six-year interdisciplinary study of the end-use of energy in a single residential community. Some results of this study are summarized in a recent book (Socolow, 1978); additional statistical analyses are presented in Mayer (1978a, 1979a), Horowitz and Mayer (1977), Mayer and Horowitz (1979) and Tittman (1978). The second is directing the Energy Information Administration-sponsored Princeton Resource Project, which is validating and improving methodologies for estimating domestic and international resources of crude oil and natural gas (e.g., Mayer, et al., 1979). The third is having directed a study for the Department of Commerce which produced a critical review of large-scale econometric energy models (Mayer, 1979b).

The perspective owes its theme to John Shewmaker, Deputy Assistant Administrator for Energy Information Validation of the Energy Information Administration, who asked me, "What does it mean for a model to be valid?" Before I could answer, he warned me that he had recently posed the question to a dozen people in the business and received 12 different answers, none of them satisfactory. Well, I thought of the old Buddhist adage that says if 12 wise men, or women, give different answers to a single question, then it must be the wrong question. In the spirit of this adage, I have concluded that the question Shewmaker posed appears simple and direct but is actually compound and complex. To answer it honestly we must develop a new perspective on the issue.

This perspective distinguishes components of the modeling process as indicated in Figure 1. The modeling of any empirical process begins with a <u>conceptual approach</u>. This approach includes a theory or a pre-theory about how the process functions and some prior expectations about the kind of evidence that could disconfirm the theory. The conceptual approach is joined with a <u>methodology</u>, a set of procedures for developing the theory into a metaphor for the empirical process. The methodology is applied to an <u>information base</u>, which includes a set of data, to produce a "<u>model</u>," a system of equations, or other analytic system, and a set of rules governing the use of that system. The term model, used in this sense, is put in quotes because the entire intellectual product is, in some sense, the model, and confusion may arise from blurring the distinction between a "model" and a model. The former is an analytic structure and a set of rules. The latter includes the former but also includes the conceptual approach, methodology and information base used to produce the "model."

The uses made of the model comprise the other half of the perspective. These include all uses, both advertised and unadvertised, those that depend directly on the forecasts of the "model" and those that depend on the existence of the model and only indirectly on its forecasts. There is a tendency among modelers to assess model use in the most esoteric fashion, as if models were used only by angels involved in rational debate over zoning the environs of heaven. Models are policy agents and political weapons and must be studied as such.

Seen through this perspective, the components of the modeling process have been confused habitually in validation studies. The very question, "What is a valid model?" tends to blur the distinctions among them and should be replaced by questions like these:

i)   What does it mean for a conceptual approach, methodology, and information base to be appropriate?

ii) What does it mean for a methodology to be optimally applied?

iii) What does it mean for a model to provide accurate forecasts?

iv) What does it mean for a model to be an effective political agent?

As these questions indicate, the issue of validity can be split into three separate problem areas. The first area deals with the validity and appropriateness of the conceptual approach, the methodology, and the information base. These are problems of science and must be treated as such. The second problem area concerns the optimality of the application of the methodology as it is used to produce the "model" and the operating characteristics of the "model" produced. These are problems of verification (dear to statisticians) including: model specification, alternative functional forms, aggre-
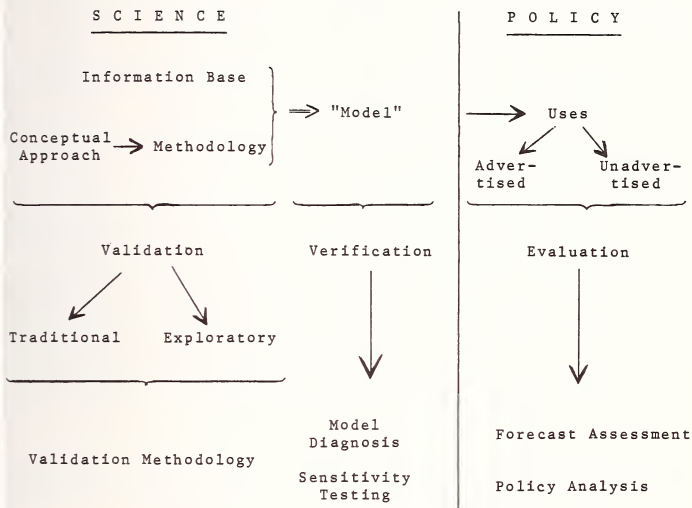
Figure 1. The Components of the Modeling Process.

gation, specification of error distributions, alternative estimation schemes, alternative forecasting mechanisms, and sensitivity analysis. Some of our colleagues tacitly define validation solely to be the tackling of these problems possibly because they, as modelers and as statisticians, feel most comfortable working in this area.

The third problem area involves the analytic study of the uses made of the broader model and its forecasts. These questions fall in the domain of policy analysis and require an approach that is scientific but cognizant of the fact that political tools, once developed, are used in ways other than those used in the justification for their creation. To ignore the variety of uses made of a model and only study the accuracy of its forecasts is both politically naive and dangerous. If we engage in energy modeling to assist policy makers and policy analysts with their problems, then we must assess the worth of our products relative to those problems. I propose a 100-year moratorium on admonishing our clients for using models in ways that appear to be less than scientific, particularly when we anticipated these uses before we undertook the model.

If the primary use of a model is to limit or expand the debate on a particular energy policy, to table or hurry consideration of a particular energy bill, to convince the public of the wisdom of a particular political position, or to focus attention on a particular energy technology at the expense of other approaches, then these are the uses that must also be assessed. It is ruinous to try to classify the uses of models as legitimate or illegitimate and then to ignore the latter. Models, like all other policy-relevant science products, are political weapons. Any good politician or bureaucrat uses these weapons to his or her political advantage. To neglect to do so might endear the politician to the modeling community but would be ineffective politically.

Within this new perspective, the question of what is a valid model is ambiguous. It could be a question about any of the components: the scientific base, the implementation of the methodology and resultant "model," or the uses made of the model. We must begin to employ the more specific questions suggested above.

Of course, the perspective proposed might not be embraced by the general public, its leaders and its press, and, in fact, is probably out of focus with any neophyte's notion of validity. I suggest that we convince these groups that a valid scientific model is produced by a valid methodology, and that a valid methodology satisfies the following property: <u>If</u> the conceptual approach is correct, and <u>if</u> the assumed scenario eventuates then, <u>on the average</u>, the forecasts generated by models produced by correct application of the methodology will be accurate and will display minimal variance from the appropriate values.

This definition does not imply that a valid scientific model will correctly predict the future. No such requirement is reasonable since science cannot see the future. We are not seers. Furthermore, this

efinition of a validity _is_ closely related to the criteria used to
valuate most physical science research efforts. Few physicists would
udge the validity of a fusion research project by assessing the re-
ults of a single experiment. There is, however, a tendency for
eople, and the press, to evaluate modeling efforts by whether a
articular model produced in the past is able to forecast the state
f nature today. Is it any wonder that many modelers, at least in
heir more Freudian recesses, fear evaluation and validation? These
riteria would scare scientists in any domain of inquiry.

Having split the problem of model validity into problem areas
nd having endorsed the importance and interrelation of each area,
would like to concentrate, henceforth, on the two extremes.

2.1 Analyzing the Approach and the Methodology

The problem of evaluating the conceptual approach, method-
logy and information base of a modeling effort is a venerated problem
n the study of the foundation of science and, accordingly, there exist
everal similar, but distinct, sets of norms by which these components
an be evaluated (Popper, 1959; Hanson, 1958; Kuhn, 1962; Hempel, 1966).

These philosophers, and others, have approached the problem of
cientific development from two standpoints. In the "traditional"
iew, the critical feature is that no theory is sacred, since the
ower of science lies in its power to disconfirm, that is, to discon-
irm theories induced from experience or from one set of data by the
eduction of test hypotheses and the checking of these against other
ata. Under the "exploratory" hypothesis of the scientific endeavor,
owever, all the crucial and creative work of science lies in the
nitial confrontation with the data, the confrontation that produces
heory.

Elsewhere I have written on some of the implications of these
iewpoints for energy analysis (Mayer, 1979c), and the broad outlines
f that work are these:

If modeling were a traditional science, the modeler would begin
ith a theory, formulate a model, use data to estimate the parameters,
est the model and thus the theory against the data, and then reformu-
ate the model and possibly the theory. Accordingly, to evaluate
odeling as a traditional scientific effort, we would have to ask
uestions like these:

1. Is the theory underlying the model explicitly presented?

2. Is the theory correctly represented in the model?

3. Are the data used appropriate for the model?

4. Is the methodology used to implement the approach appro-
    priate for the process being modeled?

5. Is the model challenged by confrontation with the data?

6. If the model fails to fit the data, are the model and theory reformulated in light of the negative evidence?

7. Does the modeler refrain from testing the reformulated model with the same data used to test the original model?

8. Are the inferences made by the modeler replicable in the sense that a second modeler beginning with the same approach and using the same methodology would reach the same conclusions?

If modeling were an exploratory science, however, we would exp
a modeler to begin his or her work with an examination of data in
search of observations worthy of investigation, then to use visual
methods to explore the data to ferret out trends and patterns in
these observations, and finally to use these trends and patterns to
suggest models worthy of further examination through traditional
studies. (For excellent introductions to the techniques of explora
tory data analysis, see Mosteller and Tukey, 1977; Tukey, 1977.)
The modeler refrains from making any statistical inferences about
any population at large. If these activities comprise scientific
modeling, then the questions we must ask include these:

1. Did the modeler refrain from adopting, or even endorsing, a model prior to examination of the data?

2. Were a rich variety of exploratory methods used or did the modeler appear to be disposed to using only a few techniques?

3. Were a wide variety of patterns sought or did the modeler have prior biases toward finding certain types of patterns?

4. Did the analysis produce patterns and relationships worthy of further investigation, possibly through traditional studies?

5. Did the modeler refrain from making any statistical inferences about the population from which the sample was drawn?

6. Would an independent exploratory effort produce patterns similar to those reported?

Surely, both lists of questions are suggestive, not exhaustive.
And just as surely, my colleagues have noted that few policy modeli
efforts are purely traditional or purely exploratory. Fortunately,
though, even if modeling lies somewhere in between traditional and
exploratory endeavors, it remains appropriate to analyze it by cer-
tain combinations of both techniques. For example, most will find

t completely appropriate to assess a model's development by questions related to the traditional approach and to examine its fit to data by questions related to the exploratory philosophy. And clearly, it is appropriate for the modeler himself to combine these two approaches in his or her work. But not every combination is legitimate, and we must be aware, for example, of the modeler who tries to develop a model by exploring a set of data and then tries to use traditional confirmatory techniques and the same set of data to test the model.

Unfortunately, certain of these illegitimate combinations are prevalent in policy modeling. For example, models are often casually developed from a convenient but vaguely formulated theory which is never made explicit. Functional forms often do not follow from theory but are chosen to facilitate interpretation and to include variables for which data are conveniently available (even if the data are obtained from sampling units other than those addressed in the theory). Parameters are estimated and hypotheses are tested from the data with copious attention paid to the statistical properties of the estimators and the tests, and if the model does not adequately "explain" the data, then the parameter estimates and hypothesis tests are used to modify the model by the minimal amount needed to fit the data. For example, a parameter will often be set to zero if the sign of the parameter violates the data or the prior expectations of the modeler. The modified model is then estimated from the same data used to estimate the original model, as if the reformulated model followed directly from theory. Hypotheses about the reformulated model are tested and the biases in the probability statements are ignored. Forecasts are generated for future values of the endogenous variables. The model is again minimally modified and re-estimated if the forecasts violate the data or the expectations of the modeler. Substantive interpretation of the effort assumes that it is the correct model of the process under study and involves interpretation of the parameter estimates, hypothesis tests, and the forecasts associated with the twice modified model.

I offer the possibly radical suggestion that the above script outlines a drama that may provide enormous amounts of useful or even accurate information about the energy world, but it is not the script of the drama of traditional or exploratory science. Violations of scientific method under the traditional view include: theory is never made explicit; theory and model are preserved regardless of the evidence contained in the data; and models are formulated, estimated, and tested from the same data. Violations of scientific methods as seen through the exploratory philosophy include: the data are never allowed to suggest a model, but only to rescue one; residuals from the estimated model or theory are not used to suggest major alterations in the model or theory; and interpretation rests solely on estimates, forecasts, and tests.

I suggest that, in fact, the above script is written for the drama of displaying theory, of using data to quantify and elaborate a set of prior beliefs about the processes being modeled. If these beliefs lead to a model which, after minor modifications, explains

the data reasonably well, then the beliefs are endorsed as being abl to produce a satisfactory model of the process at hand. Although th might be a satisfactory approach to modeling energy-related processe it is a sham to call it science. Is it any wonder that policy maker are suspicious that we incorporate our personal beliefs on important issues into our formulation of models?

Addressing a single approach to economic modeling in a recent article on the health of economics as a discipline, Lester Thurow, the distinguished economist, makes a similar point:

> Initially, econometric models were supposed to test whether the clearly specified theory could be statistically verified. Was the theory supported by the data?
>
> In the end conclusive tests did not prove to be pos- sible. Econometric models proved not to be up to the task. Equations and coefficients were not stable. Good historical equations proved to be poor predictors of the future. New data led to new coefficients.... It also proved possible to build models that were equally good statistically from a number of different perspectives. Theories could not be accepted or rejected based on the data. Equations could not stand up over time. At any moment in time the models look solid and precise, but they are in fact elastic. The data simply were not powerful enough to test and to choose among theories.
>
> As a result, econometrics shifted from being a tool for testing theories to being a tool for exhibiting theories. It became a descriptive language rather than a testing tool. Statistical models are built to show that particular theories are consistent with the data. But other theories are also consistent with the data and only occasionally can a theory be rejected because of the data. As a result good economic theory was stronger than the data -- at least in the mind of the economists -- and therefore it must be imposed on the data. What started out as being a technique for elevating data relative to theory ended up doing exactly the opposite. (Thurow, 1977)

While some of us might find his criticism a little strong, I cannot imagine that any of us don't believe that it contains a nugget of truth.

In summary, for a modeling effort to be scientific, some combina tion of two components must be present: either the theory used to justify the model must be confronted by the data, or the information base used to develop the model must be explored to uncover underlying patterns. Many energy modeling efforts attempt neither.

2.2 Analyzing the Uses Made of the Model

The problem of analyzing the uses made of the model is a problem of policy analysis whose solution begins by separating the uses of the model into advertised uses and other uses, with the intent of studying both. It is not suitable to dismiss all uses of the model other than the most esoteric as illegitimate; the challenge is to analyze models as they are employed. As scientists we may not favor the game of politics, but it _is_ the game that pays for our models.

Analyzing the advertised uses of the model is fairly straightforward, and that is one reason it receives so much attention. It usually involves assessing the accuracy of parameter estimates and forecasts relative to the needs of the major users of the model. As suggested to me by Jim Finucane of the Office of Energy Information Validation, one strategy begins by estimating how inaccurate estimates and forecasts can be before major policy decisions made by the users are affected and then concludes whether the "model" provides this needed accuracy. Econometricians and statisticians have been working for years on developing methods for addressing this issue rigorously. It is important to note that this type of assessment requires indicators of the uncertainty associated with the parameter estimates and forecasts generated by the model -- the type of information generated by confrontation with the data. The EIA has made a major improvement in currently used models by providing such indicators.

The unadvertised uses of the model are those that do not depend directly on the model's ability to predict the future. These uses may include convincing a doubting public of a particular political position, or forcing citizens without access to a large computer model out of a political debate. Moreover, these uses may be the most important in the sense that they are often the ultimate, authentic justification for funding a modeling effort.

Unfortunately, little work has been spent on the problem of assessing these unadvertised uses and few methods of assessment are available. This lack of development bespeaks the need for further research in this area; it does not remove our professional and ethical responsibility to analyze the unadvertised uses of our models, particularly since many of these uses are anticipated by the client and modeler alike before the modeling is undertaken, making the modeler a conscious though passive contributor to them. If we accept money for modeling fully or partly aware of the way that the model is likely to be used, develop a model that satisfies the political needs of our sponsor, act aghast when the model is used as a political weapon, assert that we are not responsible for the indiscretions of our sponsor, plead innocent to the charge that we serve the political process, and then search for a new sponsor so that we can begin the ruse again, then we are hypocritical and dangerous science pretenders.

We _are_ responsible for the uses made of our products. We are responsible for analyzing these uses, endorsing those uses we believe are appropriate, and collectively condemning dangerous misuse. Possibly we should all re-study the intense debate that has torn through physics for three decades regarding the proper and improper uses of atomic knowledge and the responsibility of the physics community for such uses.

Of course, there are those who argue that we should not become involved in the analysis of the uses, advertised or unadvertised, made of models as policy agents because as scientists we should stay away from politics. I will give only three responses. First, by accepting money to model processes closely related to important political debates we consent to being part of the debate. To ignore our role is to take a strong, if naive, political position. Second, the political process can be analyzed without endorsing a political position. As a beginning, one can develop a menu of the political uses made of models without falling prey to politics. Third, the analysis of the uses made of models probably does not involve biases any more serious than those contained in the conceptual approach and methodologies used by policy process modelers. Just as we need to be more explicit about the biases that go into our models, we need to be explicit about our biases regarding the legitimate use of models.

There appears to be considerable desire on the part of some policy modelers to "work both sides of the street." Although they might admit over a beer that the model they have generated reflects all of their biases about the world, when questioned in public they are not even willing to admit that they have political positions. They argue that as scientists their work is value-free and that their products represent scientific statements about reality and not any political position. I suggest that this claim has three fatal flaws. First, no science, especially policy modeling, is value-free. Science is objective only in the sense of being conditionally replicable given an approach and methodology; it is not and has never been value-free. Second, all professions that are related to policy have strong biases about the way in which the political world _should_ work. As a member of a discipline each of us operates from the normative base imprinted on us through our profession. We should not attempt to hide that base. Third, the public is not as stupid as this game requires. Let's wise up before we are all put out of business.

Every individual who steps foot in the political arena becomes a participant. The makers of weapons are part of the war.

3. APPLICATION OF THE PERSPECTIVE

As an illustration of the proposed perspective, let us assess M. King Hubbert's model of domestic crude oil production rates (see Mayer et al., 1979). Hubbert's model is a simple single-equation model and probably the most widely used tool for estimating ultimate

...estic oil production. We chose this model as an illustration be-
...se it has had tremendous impact on American energy policy and be-
...se it is distant enough from those models dear to our collective
...rts that it is unlikely to provoke defensive debate among us about
... merits. Instead, we can focus our attention on applying this
... idation perspective.

In 1956, Dr. Hubbert, a respected geologist and, at the time, an
...loyee of one of the major oil companies, estimated that ultimate
...estic (lower 48 states) oil production would amount to 170 billion
...rels, an estimate regarded as spectacular because it clashed radi-
...ly with the then-conventional wisdom of both government and industry.
...vivid was this clash that attempts were made immediately in several
...rters to discredit Hubbert's methodology.

Hubbert's conceptual approach was based on three major assump-
...ns. First, the amount of oil discovered in a year can best be de-
...ed as the sum of the amount of oil produced in that year and the
...unt of oil added to reserves in that year, a most unusual definition
..."discovery" since it is not the amount of oil contained in fields
...nd that year. Second, cumulative discovery and cumulative production
...low identical growth curves with the former leading the latter
...ough time by a constant lag of about 11 years. Third, cumulative
...covery and production are symmetric curves in time, and thus the
...e of decrease in production per year will mirror the historical
...e of increase in production per year.

With these three assumptions and the historical data available
...bert obtained his estimate of ultimate domestic production. The
...roach was original because previous estimates had relied on complex
...logical and engineering analogies or simple volumetric calculations,
... all Hubbert's method required was extrapolation of the historical
...es of discovery and production into the future under the assumptions
...t mentioned. The method yielded estimates of future discovery rates,
...ure production rates, years of peak production and peak discovery,
..., most importantly, the ultimate domestic production.

In implementing this approach, Hubbert looked at the historical
...ve of cumulative discovery -- not production -- and decided it
...ld be approximated by a simple, three-parameter logistic model
...ch he fit with a ruler by graphing the data on logarithmic paper.
... of the parameters, ultimate production, he estimated in 1956 at
... billion barrels by best professional guess and not from the data.
...s value was then plugged into the model to obtain the values of the
... remaining parameters. Not until 1962 did Hubbert attempt to
...port this estimate by direct estimation from the data, and even
...n he used an informal estimation scheme, ignoring standard statis-
...al methods such as least squares or maximum likelihood. In addition,
...did not examine the characteristics of his model through methods
...h as residual analysis or sensitivity testing.

Under the perspective we have suggested, validation of the Hubbe model begins with an evaluation of his conceptual approach, methodolc and information base as components of science. All three seem extrem ly weak, whether regarded as components of a traditional or an explor tory effort.

The very quality that made Hubbert's work so original -- its ability to ignore geological factors -- doomed it as an effort of traditional science. No fundamental theory underlay the choice of approach, methodology, or data. Hubbert's assumptions about discover and production rates, and the wisdom of using them to estimate ulti- mate production may be correct, but they were based on the most infor mal of observations and never justified. Worse still, the model's structure was never challenged by available data and thus not opened to the most important challenge of traditional science: disconfirmat Hubbert's approach may work, but it is not scientific in the traditio sense because its geological, economic and technical assumptions go unsupported and unchallenged.

Viewed as a work of exploratory science, Hubbert's approach and methodology certainly fare no better. Although Hubbert did not vio- late the criteria of exploratory science by trying to force a theory on the data, he made no effort to explore alternative hypotheses, to examine residuals, to estimate the sensitivity of his estimates to minor alterations in the model, or to search for patterns in the data The only statistics he computed were the correlation coefficient and the parameter estimates. Although Hubbert was clearly an astute ob- server of the state of the domestic oil industry, and although this keen sentience led him to forecasts that proved excellent, these fact make him a wise man but not a good scientist. He learned from exper- ience but not from the systematic processing of information on the industry.

Hubbert's implementation of his methodology was, as we have seen equally casual and his examination of the "model" obtained, nonexister He used only informal methods of parameter estimation, including an after-the-fact justification of his original 170 billion barrel pro- fessional guess, and ignored standard statistical estimation methods, analysis of residuals, and tests for sensitivity. The strategy of modeling the rate of discovery directly instead of obtaining the pro- duction model by shifting the discovery model was never considered, the sensitivity of the model to possible errors in the data was ignore

In summary, Hubbert's conceptual and methodological approaches satisfy neither the criteria of traditional science, nor those of ex- ploratory science, nor any valid combination of the two. Hubbert was satisfied with the procedure because it produced a high correlation and supported his personal judgment. Through personal communication, Hubbert has freely acknowledged that he chose the logistic model be- cause it was the only growth curve he knew about, that he had always entertained some suspicions about the worth of statistical methods in general, and that he was convinced that 170 billion barrels was the "correct" value long before he attempted any modeling.

To complete our validation of the model, we must assess the uses to which it was put. Among the advertised estimates are the years of peak discovery, reserves, and production. These estimates thus far have proved very accurate, and their success was crucial in making the incredible success of the unadvertised use of the model: the deflation of the optimistic estimates made at the time by the oil companies and the federal government.

Although immediate reaction to Hubbert's model was defensive and negative, the fact that Hubbert's advertised forecasts came true one by one was enormously effective politically. In 1974, for example, the U. S. Geological Survey revised its estimate of ultimate production to bring it more in line with Hubbert's. By 1976, the signs of success were clear: Hubbert's model had appeared in numerous scientific publications; he had been asked to contribute to three major National Academy of Science studies; he had been commissioned by the U. S. Senate to produce a committee document on the future of domestic oil production; his methodology was being used by scientists working on related problems, such as estimation of ultimate world production; and values near his were being used in almost every major energy policy document. Finally, in 1978, he won the Rockefeller Public Service Award for contributing this model to the public good.

In essence, by producing a simple model, by repeatedly publishing the same estimate of ultimate production, allegedly reached through different procedures, by correctly estimating the year of peak production and other key years, Hubbert cornered the estimation market. The complex political response of the U. S. Geological Survey to Hubbert's model provides an excellent example of the power of the model's success in advertised uses in affecting its unadvertised uses.

If, then, I were to use our venerated academic scale to assign grades to Hubbert's modeling effort, I would award him a D for approach and methodology, a D- for implementation and "model" assessment, an A for advertised uses, and an A+ for unadvertised uses. Clearly, this evaluation cannot be mapped into an answer to the simple question, how valid is his model?"

CONCLUSION

The difference between our proposed perspective and many of the prevailing views of validation can be described in terms of the substitution and clarification of issues. For the issue of whether modelers are doing a good job, we substitute two issues: first, whether modelers and their models are teaching us something about the energy world, and second, whether their products are important and effective in the process of formulating energy policy.

More specifically, for the issue of validity, this perspective substitutes three issues: first, the degree to which the model is developed by reasonable application of scientific method; second, whether the methodology employed is implemented optimally and whether

the resulting "model" displays reasonable operating characteristics; and finally, whether uses of the model, both advertised and unadvertised, are effective and appropriate for the "model."

In judging the scientific merit of the modeling effort, the validator should spend little time assessing whether the world works as assumed by the modeler. Lacking omniscience, the validator has no more chance of assessing this issue than does the modeler whose work is at issue. Instead, inquiry should focus on whether the modeler has provided scientific evidence sufficient to support his conceptual approach, methodology, and information base. It is unequivocally clear that as scientists our prior position should be that a model is inadequate until it is justified to our satisfaction. To assume otherwise may comfort a modeler but is not acceptable scientific posture. As validators we cannot be expected to discover and demonstrate which models are correct, but we can be expected to discover which models have been provided with credible scientific support. Furthermore, should all models of a particular genre lack empirical support, then it is our obligation as scientists to dismiss the entire class -- not as false, but as less than scientific.

In judging the implementation of a modeler's methodology, the validator should concentrate on developing procedures to diagnose the adequacy of various implementation routines, particularly in light of the needs of the model's user(s) and the infirmities known to exist in energy data. In addition, the validator might search for unrealistic assumptions, such as tacit assertions that all variables in the model follow a normal distribution, that any two error terms are uncorrelated, that there is no measurement error associated with the data, or that all relations are linear. In judging the operating characteristics of the "model," the validator may choose from a plethora of available methods, including those designed to judge the sensitivity of forecasts to alternative models, the sensitivity of the model to errors in the data, and the sensitivity of the model to variation in key parameters. Fortunately, these and related problems have already attracted much attention from the modeling community, attention I strongly encourage.

Finally, the validator must list and judge the advertised and unadvertised uses of a model in the political process. Advertised uses are best judged by isolating those decisions made in the policy process that could be affected by the model's forecasts, then estimating the maximum degree of variation in the forecast that would not affect the decision, and finally, comparing this degree of variation with the uncertainty associated with the forecast -- uncertainty that may be associated with the internal structure of the model or with the external fact that the model is incorrect in some fashion. Unadvertised uses of a model are more difficult to examine and assess, and thus far this process has cried in the wilderness for attention.

As modelers and as human beings, we all know that things are not always as they seem. Often -- possibly too often -- models are commissioned for reasons having little to do with those expressed explicitly. Models can be used as political weapons to a variety of purposes, including focusing attention on certain issues at the expense

others, providing a framework for envisaging the future, and enhancing or limiting debate. As policy analysts, validators must develop methods for isolating and assessing these uses. The world is too breathtaking, and models too exciting, for modelers to assume, or pretend, they are plasma physicists operating in a sterile laboratory on the dark side of the moon.

Finally, we must also recognize that a particular combination of approach, methodology, and information that carries little scientific justification -- as in the case of Hubbert's model -- could produce the finest of forecasting tools. As scientists, we should condemn such a model as inadequately supported, but we must make the public aware of why this is a crucial statement. No scientist can claim that tools created by an other-than-scientific method will be less accurate than those developed by rigorous and vigorous application of the scientific method, but as scientists we do believe that the creation of energy models by application of the scientific method is an important part of energy analysis, and that, in the long run, the policy process will be better off if legitimate scientific models are generated.

DISCUSSION

Dr. Everett (DOE): One fact, domestic crude oil production peaked in 1970, so Hubbert was very close.

Dr. Mayer: Right.

Dr. Everett: Two, he must have cornered the market and got a very high grade in the policy process because he wouldn't come to lunch when we asked him to come!

REFERENCES

[1]   Hanson, Norwood R. (1958). Patterns of Discovery. Cambridge.

[2]   Hempel, Carl G. (1966). Philosophy of Natural Science. Prentice-Hall.

[3]   Horowitz, Cynthia E. and Mayer, Lawrence S. (1977). The Relation ship between the Price and Demand for Natural Gas A Partially Controlled Study. Energy Research, 1 193-222.

[4]   Kuhn, Thomas S. (1962). The Structure of Scientific Revolutions University of Chicago Press.

[5]   Mayer, Lawrence S. (1978a). Estimating the Effect of the Onset of the Energy Crisis on Residential Energy Demand Resources and Energy, 1, 57-92.

[6]   _____ (1978b). Exploratory Data Analysis and Class ical Statistics: Their Ability to Shed Light on Energy Issues. Department of Energy 1977 Statis tical Symposium Proceedings, 27-32.

[7]   _____ (1979a). The Use of Semi-Controlled Experi ments in the Analysis of Residential Energy Deman Department of Energy 1978 Statistical Symposium Proceedings (forthcoming).

[8]   _____ (1979b). Econometric Energy Models: A Crit ical Review. Technical Report, Department of Statistics, Princeton University.

[9]   _____ (1979c). The Use of Exploratory Methods in Economic Analysis: Analyzing Residential Energy Demand. To appear in Criteria for Evaluating Econometric Models, J. Kmenta and J. Ramsey, Eds.

[10] Mayer, Lawrence S. and Benjamini, Yoav (1978). Modeling Residen tial Demand for Natural Gas as a Function of the Coldness of the Month. In Saving Energy in the Home, R. Socolow (Ed.), Ballinger, 301-312.

[11] Mayer, Lawrence S. and Horowitz, Cynthia E. (1979). The Effect of Price on the Residential Demand for Electricit A Statistical Study. Energy, 4, 87-99.

[12] Mayer, Lawrence S., Silverman, Bernard, Zeger, Scott L., and Bruce, Andrew G. (1979). Modeling the Rates of Domestic Crude Oil Discovery and Production. Technical Report, Resource Project, Department of Statistics, Princeton University.

[13] Mosteller, Frederick, and Tukey, John W. (1977). Data Analysis
              and Regression. Addison-Wesley.

[14] Popper, Karl R. (1959). The Logic of Scientific Discovery.
              Harper and Row.

[15] Socolow, Robert S. (Ed.) (1978). Saving Energy in the Home:
              Princeton's Experiments at Twin Rivers. Ballinger.

[16] Thurow, Lester (1977). Economics: 1977. Daedalus, 11, 79-94.

[17] Tittman, Peter C. (1978). Conservation Trends in Natural Gas
              Consumption from 1974 to 1978. Unpublished
              report, Department of Statistics, Princeton Uni-
              versity.

[18] Tukey, John W. (1977). Exploratory Data Analysis. Addison-
              Wesley.

SYSTEMATIC SENSITIVITY ANALYSIS
USING DESCRIBING FUNCTIONS

Fred C. Schweppe and James Gruhl

M.I.T. Energy Lab, Cambridge, Mass. 02139

1. INTRODUCTION

This paper discusses simple, straightforward procedures for performing systematic sensitivity analysis on the input-output behavior of large mathematical models that are implemented as digital computer programs. The techniques were developed as part of an EPRI-funded study undertaken at the MIT Energy Laboratory, Model Assessment Group on the validation of the Baughman-Joskow "Regionalized Electricity Model" (REM); see, for example, reference 1. REM is used only as an example to give focus to the general ideas, and no conclusions or results about REM itself are given (these can be found in other documents such as in reference 2).

REM is a sophisticated computer program that simulates the dynamic behavior of portions of the U.S. energy supply/demand market with particular emphasis on the electric sector. For the sake of developing a simplified mathematical representation of REM define:

$\underline{\alpha}$: vector of exogenous input parameters, and

$\underline{y}$: vector of model outputs.

The elements of $\underline{\alpha}$ and $\underline{y}$ can be generalized to cover series of discrete, or even continous, functions in time. The concepts of this paper, however, were developed and tested using the simple constant $\underline{\alpha}$ and $\underline{y}$ as defined in Figure 1. The $\underline{y}$ are outputs in 1997 which is the terminal year for the base case simulation.

Once the inputs $\underline{\alpha}$ and outputs $\underline{y}$ have been defined, any large computer model can be viewed simply as a nonlinear function $\underline{f}$ which translates the $\underline{\alpha}$ into the $\underline{y}$:

$$\underline{y} = \underline{f}(\underline{\alpha}) \tag{1.1}$$

$\underline{\alpha}$: vector of exogenous input parameters

$\underline{y}$: vector of model outputs in 1997

This point of view is valid independent of whether the model is dynamic or static, a simulation or an optimization, deterministic or stochastic, etc.

To address the issue of sensitivity analysis of a model, define:

$\underline{\alpha}_0$: base case input

$\underline{y}_0 = \underline{f}(\underline{\alpha}_0)$: base case output

$\Delta\underline{\alpha}$: input perturbation (e.g., uncertainty)

$\underline{\alpha} = \underline{\alpha}_0 + \Delta\underline{\alpha}$: perturbed input

$\underline{y}(\Delta\underline{\alpha}) = \underline{f}(\underline{\alpha}_0 + \Delta\underline{\alpha})$: perturbed output

$\delta\underline{y}(\Delta\underline{\alpha}) = \underline{y}(\Delta\underline{\alpha}) - \underline{y}_0 = \underline{f}(\underline{\alpha}_0 + \Delta\underline{\alpha}) - \underline{f}(\underline{\alpha}_0)$

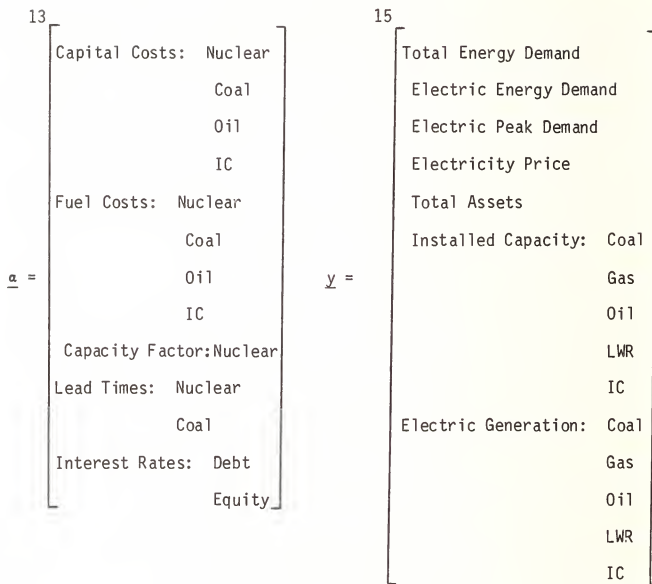$= $ output perturbation (e.g., uncertainty).

Figure 1

REM Inputs and Outputs Used in the Following Examples

498

A vast amount of material can be gathered for model sensitivity analysis if the following question can be answered:

> Given a characterization of $\Delta\underline{\alpha}$
> what are properties of $\delta\underline{y}(\Delta\underline{\alpha})$? (1.2)

In fact one might define the whole process of sensitivity analysis as the understanding, testing, and evaluation of the properties of $\delta\underline{y}(\Delta\underline{\alpha})$.

The very nature of $\delta\underline{y}(\Delta\underline{\alpha})$ suggests that it might be developed by a differentiating process, and in fact a Taylor series expansion of (1.1) yields

$$\delta\underline{y}(\Delta\underline{\alpha}) = \underline{J}\,\Delta\underline{\alpha} + \text{higher order terms of}\,\Delta\underline{\alpha},$$

where

$$\underline{J}: \text{Jacobian} = \left.\frac{\partial\underline{f}(\underline{\alpha})}{\partial\underline{\alpha}}\right|_{\underline{\alpha}=\underline{\alpha}_0} \qquad (1.3)$$

If the higher-order terms in (1.3) could be ignored, the Jacobian $\underline{J}$ itself could provide the answer to (1.2). Unfortunately, for many cases of concern, the nonlinearities represented by the higher order terms are critical. For the REM example the Jacobian $\underline{J}$ was a 13 by 15 matrix and it was numerically estimated three different ways:

(1) using a 1% perturbation in $\underline{\alpha}$, which is partially displayed in Table 1,

(2) using a 10% perturbation in $\underline{\alpha}$, partially displayed in Table 2, and finally

(3) using a 20% perturbation in $\underline{\alpha}$, displayed in Table 3.

The fact that these three different numerical estimates were radically different in many of their entries demonstrated that the model's behavior was significantly nonlinear, thus question (1.2) could not be adequately answered by simple linear characterizations.

Two approaches for dealing with this inherent nonlinear nature of the model were developed:

(1) Criterion Sensitivity Analysis, and

(2) Describing Functions.

This paper emphasizes the Describing Function approach but the Criterion Sensitivity Analysis concept is briefly reviewed in the next section.

2. CRITERION SENSITIVITY ANALYSIS

Some of the basic ideas underlying the Criterion Sensitivity Analysis approach are reviewed here in a simplified fashion. More details can be found in Ref. 3.

Again consider the question posed in (1.2). Assume that the characterization of the input perturbation (uncertainty) $\Delta\underline{\alpha}$ is expressed as

INSTALLED CAPACITY IN 1997

		COAL	GAS	OIL	LWR	INT.COM.
	NUCLEAR	1.5083	-0.0004	0.0012	-1.4855	-3.7630
CAPITAL COSTS	COAL	-0.7900	-0.0004	-0.0005	0.8967	-0.0028
	OIL	-0.0136	-0.0004	-0.0021	-0.0025	0.0390
	INT.COM.	-0.0009	-0.0004	-0.0005	-0.0099	-0.0209

Table 1   Capital Cost and Capacity Expansion Portions of Normalized Gradients from 1% Parameter Perturbations in REM

INSTALLED CAPACITY IN 1997

		COAL	GAS	OIL	LWR	INT.COM.
	NUCLEAR	0.3648	0.0000	0.0000	-0.7008	-0.9493
CAPITAL COSTS	COAL	-0.8692	0.0000	0.4451	0.8343	-0.4042
	OIL	0.0979	-0.0000	-1.1935	0.0494	-0.2247
	INT.COM.	0.0006	-0.0000	-0.0000	-0.0095	-0.0136

Table 2   Capital Cost and Capacity Expansion Portions of Normalized Gradients from 10% Perturbations

Table 3 — Complete Table of Normalized Gradients from 20% Parameter Perturbation Results

Category	Parameter	Total Energy Demand	Electric Energy Demand	Peak Electric Demand	Average Price	Total Assets	Coal Installed Capacity	Gas Installed Capacity	Oil Installed Capacity	LWR Installed Capacity	IC Installed Capacity	Coal Electric Generation	Gas Electric Generation	Oil Electric Generation	LWR Electric Generation	IC Electric Generation
Capital Costs	Nuclear	0.0054	-0.193	-0.1816	0.2249	-0.1449	0.8556	0.0000	0.0000	-1.2579	-0.2641	1.5920	3.7776	0.4617	-1.3770	0.0079
	Coal	0.0184	-0.145	-0.1534	0.0341	0.0291	-1.6625	0.0000	9.2815	0.5137	0.0947	-1.4504	7.1722	0.8915	0.5122	-0.3031
	Oil	0.0035	0.016	0.0148	-0.0286	-0.0091	0.0702	-0.0000	-0.7747	0.0249	-0.1183	0.0208	-1.1735	0.4824	0.0265	0.2462
	IC	0.0003	-0.002	-0.0016	0.0069	0.0123	0.0007	-0.0000	-0.0000	-0.0075	-0.0209	0.0080	-0.0003	-0.1198	-0.0073	-0.0186
Fuel Costs	Nuclear	0.0066	-0.062	-0.0594	0.0667	-0.1397	0.2187	-0.0000	-0.0046	-0.4167	-0.0682	0.4442	0.0004	1.5999	-0.3992	0.1961
	Coal	-0.0179	-0.071	-0.0757	0.0915	0.0869	-1.1959	0.0000	0.2258	4.1225	-0.0254	-1.9604	-0.0635	2.1778	1.0445	0.2993
	Oil	-0.0572	0.010	0.0109	0.0095	0.0147	0.0204	0.0000	-0.0450	-0.0402	0.1317	0.1015	-0.1065	-0.2271	-0.0392	-0.0490
	Gas	-0.1782	0.157	0.1587	0.0141	0.1357	0.3494	0.0000	-0.0011	0.1373	-0.6595	-0.2235	-0.2235	1.2253	0.1402	0.2726
Rat. Utilization Factor	Nuclear	-0.0043	0.390	0.3680	-0.4162	0.8438	-2.0061	-0.0077	-0.1639	2.7315	1.1430	-3.1007	-9.3039	-13.3509	2.8677	-1.3063
Lead Times	Nuclear	0.0023	-0.085	-0.0036	0.0066	-0.0536	0.1422	0.1234	0.7453	-0.6615	0.4196	0.5965	5.1119	4.7744	-0.5449	0.1459
	Coal	0.0003	-0.009	-0.0090	-0.0062	-0.0011	-0.0365	-0.0984	-0.0237	0.0141	0.2729	-0.0380	-0.2654	-0.3347	0.0129	-0.0070
Interest Rates	Debt	0.0044	-0.157	-0.1542	0.1356	-0.2711	-0.1171	0.0000	0.0000	-0.2242	-0.5412	-0.1249	0.5491	-0.3148	-0.2141	-0.1955
	Equity	0.0048	-0.203	-0.2046	0.1474	-0.2929	-0.1234	0.0000	0.0000	-0.2233	-0.4147	-0.2114	0.1998	-0.0991	-0.2417	-0.2141

Table 3 Complete Table of Normalized Gradients from 20% Parameter Perturbation Results

$\Delta\underline{\alpha}$ is constrained to lie within some set Ω_Δ, i.e.,

$$\Delta\underline{\alpha} \in \Omega_\Delta .$$

For example, assume $\underline{\alpha}$ has two elements, α_1 and α_2. One might characterize the uncertainty in one's knowledge about $\underline{\alpha}$ by saying that, for example, "α_1 is known to within 10%" while "α_2 is known to within 20%."

In this case the set Ω_Δ is a rectangular box in two dimensions centered around zero with sides plus and minus 10% of the base case magnitude in the α_1 dimension and with "20% sides" in the α_2 dimension. Assume that some particular scalar criterion function $c(\Delta\underline{\alpha})$ of the output perturbation

$$c(\Delta\underline{\alpha}) = c\left[\delta\underline{y}(\Delta\underline{\alpha})\right]$$

has been selected. For the REM studies, the criterion function $c(\Delta\underline{\alpha})$ was chosen to be one of the 15 elements of the perturbed outputs $\delta\underline{y}(\Delta\underline{\alpha})$.

The basic idea of Criterion Sensitivity Analysis is to:

"Find the particular $\Delta\underline{\alpha} \in \Omega_\Delta$ which maximizes $c(\Delta\underline{\alpha})$." (2.1)

This can be viewed as a "worst case" sensitivity analysis which finds the particular perturbation $\Delta\underline{\alpha}$ that yields the maximum sensitivity as measured by the size of c. It would be difficult to choose a criterion $c(\Delta\underline{\alpha})$ except for the fact that one need not be limited to a single criterion. Thus if 15 c's are chosen for REM corresponding to the 15 outputs $\delta\underline{y}$, 15 different $\Delta\underline{\alpha}$ perturbations would be found, each one representing the "worst case" for the corresponding output.

Criterion Sensitivity Analysis can provide valuable insight into the behavior of a model. In particular it can show how precisely the inputs must be known to have various confidences in the output values. In the case of REM, it showed that a relatively "small" perturbation set Ω_Δ contains $\Delta\underline{\alpha}$'s which can cause massive changes in the output $\underline{y}$. It could be argued that it is "unfair" to use the size of a worst-case input perturbation as a true measure of a model's sensitivity. If, however, we recall that the real purpose for developing awareness of sensitivity is for understanding, testing, and evaluation then it seems obvious that such a process should commence with the extreme cases.

Some authors have discussed the sensitivity of models by saying, for example, "the output is insensitive to 10% changes in the input parameters," where they are making an inherent, but often not explicit, assumption that the inputs are changed one at a time. For nearly linear models this assumption is not troublesome. It is not difficult, however, to think of nonlinear situations where such an assumption could be grossly misleading. Consider for example the models:

$$y = \alpha_1 \alpha_2 , \text{ or} \tag{2.2}$$

$$y = \alpha_1^{\alpha_2} \tag{2.3}$$

where the base case is $\alpha_1 = 0, \alpha_2 = 0$. In these cases y is completely insensitive to one-at-a-time changes in α_1 or α_2, but highly sensitive to multiple changes. Criterion Sensitivity Analysis emphasizes the need for more careful discussion and definition of terms particularly in the study of highly nonlinear models.

3. DESCRIBING FUNCTIONS: THEORY AND MOTIVATION

The easiest way to understand describing functions is to consider the case of a scalar α and a scalar y. Since the general vector case is presented here, any reader who has trouble following the notation should "read" the vectors as simple scalars in the following equations.

With the describing function approach, the input perturbation vector $\Delta\underline{\alpha}$ is viewed as a random vector which is characterized by its probability density:

$p(\Delta\underline{\alpha})$: probability density of $\Delta\underline{\alpha}$

The concept allows for any probability density: uniform, triangular, Gaussian, and so on. Furthermore, it is not assumed that the individual elements of $\Delta\underline{\alpha}$ are independent.

The describing function $\underline{D}(\Delta\underline{\alpha})$ is here defined to be a vector polynomial function of the vector $\Delta\underline{\alpha}$ (although in general it could be any set of $\Delta\underline{\alpha}$ functions, in particular it might include a set of homogenous response functions):

$$\underline{D}(\Delta\underline{\alpha}) = \underline{A}_o + \underline{A}_1\Delta\underline{\alpha} + \sum_{m=1}^{M} \underline{e}_m\Delta\underline{\alpha}^T \underline{A}_{2m}\Delta\underline{\alpha} + \text{cubic terms} + \ldots \qquad (3.1)$$

where M is the dimension of $\underline{D}(\Delta\underline{\alpha})$ (which is same as dimension of $\underline{y}$) and $\underline{e}_m$ is the unit column vector (all zero except for 1 in the mth row). To simplify the appearances of equation (3.1) and the following derivations, define:

> $\underline{\phi}(\Delta\underline{\alpha})$: vector of ones, $\Delta\underline{\alpha}$, and powers and cross products of $\Delta\underline{\alpha}$ up to the number of terms desired,

> $\underline{A}$: matrix of $\underline{A}_o$, $\underline{A}_1$, $\underline{A}_{2m}$ ... of (3.1), and

> K: dimension of $\underline{\phi}(\Delta\underline{\alpha})$.

where the $\underline{\phi}(\Delta\underline{\alpha})$ and $\underline{A}$ are such that (3.1) can be rewritten as

$$\underline{D}(\Delta\underline{\alpha}) = \underline{A}\,\underline{\phi}(\Delta\underline{\alpha}). \qquad (3.2)$$

The describing function problem is then defined as follows:

> Find values for matrix $\underline{A}$ and the number of terms K such that

$$\delta\underline{y}(\Delta\underline{\alpha}) \approx \underline{D}(\Delta\underline{\alpha}) = \underline{A}\,\underline{\phi}(\Delta\underline{\alpha}) \qquad (3.3)$$

503

Assume the model has been run N+1 times, that is, the base case $\underline{\alpha}_0$ plus N cases of input perturbations $\Delta\underline{\alpha}_n$, to yield N output perturbations $\delta\underline{y}_n$, $n = 1$ ... N. Now it is necessary to set up a series:

$\widetilde{p}(\Delta\underline{\alpha}_n)$: Integral of $p(\Delta\underline{\alpha})$ over "area around" $\Delta\underline{\alpha}_n$,

such that

$$\sum_{n=1}^{N} \widetilde{p}(\Delta\underline{\alpha})_n = 1. \tag{3.4}$$

The $\widetilde{p}(\Delta\underline{\alpha}_n)$ $n = 1...N$ constitutes a discretization of the density $p(\Delta\underline{\alpha})$ relative to the $\Delta\underline{\alpha}_n$ $n = 1 ... N$.

Next define

$$\underline{\xi}(\underline{A}) = \sum_{n=1}^{N} \left[\delta\underline{y}(\Delta\underline{\alpha}_n) - \underline{A}\,\underline{\phi}(\Delta\underline{\alpha}_n)\right]\left[\delta\underline{y}(\Delta\underline{\alpha}_n) - \underline{A}\underline{\phi}(\Delta\underline{\alpha}_n)\right]^T \widetilde{p}(\Delta\underline{\alpha}_n) \tag{3.5}$$

It can be shown by the usual "weighted least squares" minimization arguments that the positive semi-definite matrix $\underline{\xi}(\underline{A})$ is minimized by $\widehat{\underline{A}}$ if

$$\widehat{\underline{A}} = \left[\sum_{n=1}^{N} \delta\underline{y}(\Delta\underline{\alpha}_n)\,\underline{\phi}^T(\Delta\underline{\alpha}_n)\,\widetilde{p}(\Delta\underline{\alpha}_n)\right]\underline{Z}^{-1}, \tag{3.6}$$

where

$$\underline{Z} = \sum_{n=1}^{N} \underline{\phi}(\Delta\underline{\alpha}_n)\,\underline{\phi}^T(\Delta\underline{\alpha}_n)\,p(\Delta\underline{\alpha}_n); \tag{3.7}$$

and furthermore that

$$\underline{\xi}(\widehat{\underline{A}}) = \underline{C}_1 - \underline{C}_2, \tag{3.8}$$

where

$$\underline{C}_1 = \sum_{n=1}^{N} \delta\underline{y}(\Delta\underline{\alpha}_n)\,\delta\underline{y}^T(\Delta\underline{\alpha}_n)\,\widetilde{p}(\Delta\underline{\alpha}_n), \tag{3.9}$$

and

$$\underline{C}_2 = \widehat{\underline{A}}\,\underline{Z}\,\widehat{\underline{A}}^T \tag{3.10}$$

$$= \sum_{n=1}^{N} \underline{D}(\Delta\underline{\alpha}_n)\,\underline{D}^T(\Delta\underline{\alpha}_n)\,\widetilde{p}(\Delta\underline{\alpha}_n),$$

504

where

$$D(\Delta\underline{\alpha}_n) = \widehat{\underline{A}}\underline{\phi}(\Delta\underline{\alpha}_n).\tag{3.11}$$

Equation 3.6 provides one way of computing a set of values for $\widehat{\underline{A}}$. The motivation for the selection of this particular choice of weighted least squares is as follows. Assume the number of perturbations N is large and the $\Delta\underline{\alpha}_n$ $n = 1 \dots N$ effectively "span the space" where $p(\Delta\underline{\alpha})$ is non-zero. Then as an approximation, the various summations over n can be replaced by integrals and (3.5) becomes:

$$\underline{\xi}(\underline{A}) \approx E\left\{\left[\delta\underline{y}(\Delta\underline{\alpha}) - \underline{A}\,\underline{\phi}(\Delta\underline{\alpha})\right]\left[\delta\underline{y}(\Delta\underline{\alpha}) - \underline{A}\underline{\phi}(\Delta\underline{\alpha})\right]^T\right\}\tag{3.12}$$

while (3.8) to (3.10) become:

$$\underline{\xi}(\widehat{\underline{A}}) = \underline{C}_1 - \underline{C}_2,$$
$$\underline{C}_1 = E\left\{\delta\underline{y}\ (\Delta\underline{\alpha})\ \delta\underline{y}^T(\Delta\underline{\alpha})\right\}\tag{3.13}$$

: Mean square of output perturbation $\delta\underline{y}(\Delta\underline{\alpha})$, and

$$\underline{C}_2 = E\left\{\underline{D}\ (\Delta\underline{\alpha})\ \underline{D}^T(\Delta\underline{\alpha})\right\}$$

: Mean square of Describing Function,

where the expectation "E" is over the probability density $p(\Delta\underline{\alpha})$ of $\Delta\underline{\alpha}$. Thus the $\widehat{\underline{A}}$ of (3.6) can be interpreted as approximating the value which minimizes the mean square error as defined by (3.12).

The basic describing function problem of (3.3) also requires the determination of the number of terms, K, (degree of polynomials, number of cross products, etc.) to be used. The motivation for the logical determination of K follows from (3.13). The value of K is simply increased until

$$\underline{C}_2 \gg \underline{\xi}(\widehat{\underline{A}})\tag{3.14}$$

so that

$$\underline{C}_2 \approx \underline{C}_1\tag{3.15}$$

Thus a value of K is chosen such that the mean square of the describing function $\underline{C}_2$ is a good approximation to the mean square of the actual output perturbation $\underline{C}_1$.

In order to implement the preceding it is necessary to compute the matrix $\underline{Z}$ of (3.7). Numerically this can be done as suggested in (3.7) or one can use the integral version:

$$\underline{Z} = \int \underline{\phi}(\Delta\underline{\alpha})\underline{\phi}^T(\Delta\underline{\alpha})\ p(\Delta\underline{\alpha})\ d(\Delta\underline{\alpha})\tag{3.16}$$

This integral can be evaluated in closed form for a wide variety of $p(\Delta\underline{\alpha})$ (uniform, triangular, Gaussian, etc.) because the elements of $\underline{\phi}(\Delta\underline{\alpha})$ are polynomials in $\Delta\underline{\alpha}$.

505

The overall procedure for determining the describing function (i.e. solving (3.3)) is summarized in Figure 2. Two iterative loops are shown. The "inner loop" involves varying the number, K, and type of polynomial terms which form $\phi(\Delta\alpha)$ until the sum of squared residual $\underline{\xi}(\underline{A})$ is satisfactorily small. The outer loop involves increasing the number of input - output perturbations N.

The outer loop on Figure 2 would not be needed if the model were "cheap enough" to run so that initially N could be made very large. However with many models, the computer costs incurred per run of the model will be significant and it will be desirable to minimize the number of runs. Hence the initial choice of $\Delta\alpha_n$ to span $p(\Delta\alpha)$ may be sparse and extra perturbations may be desired in the directions that are exhibiting nonlinear or interesting behaviors. Often the choice of $\Delta\alpha_n$ will be influenced by human judgment concerning the "importance" of various issues. One obviously necessary condition is that N (number of perturbations) be "appreciably greater" than K (number of terms). In practice this type of condition should be checked in each of the individual directions of $\Delta\underline{\alpha}$.

In summary, the preceding process involves fitting a set of describing functions to a set of model runs that represent a set of points on the model's input-output response surface. The suggested manner of determining the best fit has been a weighted least squares approach, where the importances of the different points are weighted by the likelihood that the response is going to be in the neighborhood of those points. With a large enough set of describing functions, that is as K approaches N, the fitting of the different points can be "perfect." Such "perfect" fits are highly susceptible to capitalization on chance effects, and the preceding discussion suggests that one way of avoiding such spurious fits is to restrict the <u>number</u> of describing functions in the set being used. There are endless variations on this particular suggested procedure for determining the "best" fit. A few of these variations include:

(1)   minimizing the maximum residuals between the surface and the points,

(2)   minimizing absolute differences or some other robust measure rather than least squares, or

(3)   using weighted maximum residuals or weighted robust techniques.

In addition it would be possible to exploit any intuitions one might have about the true characteristics of the model's response surface. For example, if a large number of points were available and one's intuition suggested that the response surface should be a relatively smooth connection of those points, then the fitting criterion might be to minimize the integral of the deviation between the fitted surface and the piecewise linear connection of the available points with their nearest neighbors (or the supporting hyperplane n-gons connected over all convex sets of available points). The principal drawback to these more elegant techniques is that they may not be nearly as easily solved as is the least squares approach.
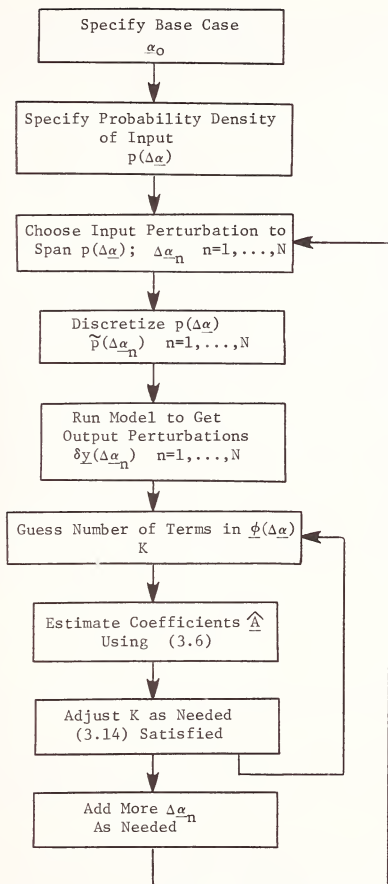
Figure 2 Flowchart Showing the Adaptive Technique for Determination of Describing Functions

The basic technique of Figure 2 is called a "describing function" approach because that term refers to certain types of closely related engineering studies which motivated the present development. It is almost a certainty, however, that the same basic ideas are used in other fields and disciplines with undoubtedly different names.

4. USES OF DESCRIBING FUNCTIONS

Assume all the steps of Figure 2 have been successfully completed and an $\widehat{A}$ has been found such that

$$\delta \underline{y}(\Delta \underline{\alpha}) \approx \underline{D}(\Delta \underline{\alpha}) = \widehat{A} \underline{\phi}(\Delta \underline{\alpha})$$

The next topic is the use of $\underline{D}(\Delta \underline{\alpha})$.

A first, and very important, step is to "plot" $\underline{D}(\Delta \underline{\alpha})$ versus $\Delta \underline{\alpha}$ in order to get a feel for the response surface of the model. Obviously one usually cannot plot the full surface at once, but various "slices" through the surface can be studied. It is extremely helpful to have a good software display system to facilitate this process of getting an understanding of the nature of the surface. Such a plotting facility can also be very useful during the iterations on K and N associated with Figure 2.

Plotting is a very helpful method of appealing to the understanding and intuition of the analyst. More explicit numerical summaries are, however, also of value to the analyst and have the advantage of being easier to use in automated procedures and in documentations. One explicit and useful numerical summary is simply the mean value of $\underline{D}(\Delta \underline{\alpha})$:

$$E\left\{\underline{D}(\Delta \underline{\alpha})\right\} = \widehat{A} \int \underline{\phi}(\Delta \underline{\alpha}) \, p(\Delta \underline{\alpha}) \, d(\Delta \underline{\alpha}) \qquad (4.1)$$

Note that even if $p(\Delta \underline{\alpha})$ is a zero mean distribution, $\underline{D}(\Delta \underline{\alpha})$ will usually have a non-zero mean because of the model's nonlinearities.

Perhaps the single most useful, explicit output is the mean square matrix of $\underline{D}(\Delta \underline{\alpha})$, i.e.:

$$E\left\{\underline{D}(\Delta \underline{\alpha}) \underline{D}^T(\Delta \underline{\alpha})\right\} = \widehat{A} \, \underline{Z} \, \widehat{A}^T \qquad (4.2)$$

where $\underline{Z}$ is given by the integral form of (3.16). By substituting the effect of the mean, equation 4.2 can be converted to a covariance matrix if desired. Such matrixes systematically summarize, on one sheet of paper in a very simple, easily interpreted format, a tremendous amount of information about the sensitivity of the model.

Specification of the first and second moments does not exactly determine the probability density of $\underline{D}(\Delta \underline{\alpha})$ as, in general, $\underline{D}(\Delta \underline{\alpha})$ will not be Gaussian. However for cases where the dimension of $\underline{\alpha}$ is large, as those familiar with a Central Limit Theorem will recall, an assumption that $\underline{D}(\Delta \underline{\alpha})$ is Gaussian may be justified, at least as a first approximation. When justified, this enables powerful statements to be made using (4.1) and (4.2).

In many situations, models are used to determine the <u>relative</u> effects of different types of hypothesized inputs, with the absolute value of the output not being as much of concern as the direction and magnitude of the changes in those outputs. Describing functions can be extremely useful in such cases. For example, suppose the model is being used to determine the relative impact of one input parameter (say α_1) on the output.

Computation of the mean and covariance matrix of the derivative $\partial \underline{D}(\Delta\alpha)/\partial\alpha_1$ will show how sensitive the model is to input uncertainties. Such formulas are easy to evaluate once the describing function itself is available.

Before beginning the concluding discussion, a couple more examples from the REM are presented. The first set of outcomes is an example of the propagation of input variation through linear simplifications of the model to get some measures of output uncertainty. Coal and nuclear plant investment costs were assigned deviations of 32%, with oil plants and internal combustors with 20% deviations, along with various deviations for the other inputs and parameters. Table 4 shows a portion of the output covariance matrix. When these variances are converted to deviations, and compared with the diagonal elements generated using the other linear models (that is, those developed from the 1% and 10% perturbations) the projections shown in Table 5. are developed. An examination of this table makes it quickly apparent once again that the REM is very nonlinear in the region around the base case. Oil-fired boiler power plants show heightened sensitivity away from the base case, while internal combustion devices show the opposite behavior. Such nonlinear behavior suggests a real need for a nonlinear simplification, such as a describing function approach.

A simple example of the use of describing functions is presented for this same investment cost and capacity expansion portion of REM. Representing $\underline{y}$ as the output vector of plant type capacities, $\underline{y}_o$ as the base case output, $\underline{u}$ as the vector of input investment costs with $\underline{u}_o$ as the base case input, then define

$$\Delta\underline{y} = \underline{y} - \underline{y}_o, \tag{4.3}$$

$$\Delta\underline{u} = \underline{u} - \underline{u}_o, \tag{4.4}$$

Using a quadratic-level describing function and a least-squares fitting criterion, then the normalized representation of this portion of the model's response is

$$\frac{\Delta\underline{y}}{\underline{y}_o} = \underline{c}_1\left(\frac{\Delta\underline{u}}{\underline{u}_o}\right) + \underline{c}_2\left(\frac{\Delta\underline{u}}{\underline{u}_o}\right)^2, \tag{4.5}$$

INSTALLED CAPACITY IN 1997

		COAL	GAS	OIL	LWR	INT.COM.
	COAL	0.3110	0.0022	-0.7355	-0.2338	-0.1365
INSTALLED	GAS	0.0022	0.0022	0.0042	-0.0047	-0.0026
CAPACITY	OIL	-0.7355	0.0042	7.8433	-0.5701	-0.2016
	LWR	-0.2338	-0.0047	-0.5701	0.5000	0.1110
	INT.COM.	-0.1365	-0.0026	-0.2016	0.1110	0.3225

Table 4 Capacity Portion of the Output Covariance Matrix for Expected Input Uncertainty Calculated from 20% Model.

INSTALLED CAPACITY DEVIATIONS IN PERCENTAGES

	COAL	GAS	OIL	LWR	INT.COM.
1% Model	79%	4%	18%	52%	265%
10% Model	48%	4%	22%	60%	75%
20% Model	56%	5%	280%	71%	57%

Table 5 Output Deviations for Assumed Input Deviations and Various Linearized Models

where C_1 and C_2 are shown in Tables 6 and 7. Although this is not a good example for demonstrating the use of describing functions for providing simplified insights, it is a good example for the development of functions that are appropriate for use in determining output uncertainty measures, and is even useful for simple out-of-model exercises. For example, if coal and nuclear plant prices <u>both</u> go up by 32% then equation (4.5) can be used to predict that coal capacity drops 11% and nuclear capacity goes down 16% relative to the unperturbed case.

5. DISCUSSION

This paper has presented approaches for performing sensitivity analysis on the input-output behavior of large mathematical models that are implemented on digital computers. Because simple linearization is not appropriate for many models of interest, emphasis has been given to techniques which can handle nonlinear behavior. The basic problem has been stated:

o Given a characterization of the input perturbation, what are the properties of the output perturbation?

Two approaches have been discussed:

o <u>Criterion Sensitivity Analysis</u>: Assumes input and perturbation are bounded (lie within a specified set); Determines maximum sensitivities (output perturbation);

o <u>Describing Function</u>: Assumes input perturbation are random variables with specified probability densities; Fits functional models over the range of input uncertainties; Determines first and second order moments of resulting output uncertainties (perturbations).

These two approaches involve basically different philosophies. Both may be worth considering in specific situations.

Sensitivity analysis is sometimes viewed as part of the "validation" of a model. However sensitivity analysis of the type discussed here inherently assumes that the model's structure is correct. Thus a model could be completely "invalid" without sensitivity analysis giving any indication of the existence of the "invalidity".

Nevertheless, the importance of doing the type of sensitivity analysis discussed here cannot be overemphasized. In almost all mathematical models, there is uncertainty in the values of the input parameters. It is usually essential to understand how these uncertainties propagate through the model and are reflected in terms of output uncertainties. Large uncertainties in the output do not necessarily invalidate the model but it is essential that the nature of these uncertainties be understood. If the uncertainties are ignored, there is a real danger that important points in the assessment of a model will be overlooked.

INSTALLED CAPACITY IN 1997

	COAL	GAS	OIL	LWR	INT.COM.
NUCLEAR	0.0758	-0.0000	0.0000	-0.3592	-1.4700
COAL	-0.3509	-0.0000	-5.3166	1.0445	-0.7181
OIL	0.1130	-0.0001	-1.4342	0.0640	-0.2849
INT.COM.	0.0004	-0.0001	-0.0001	-0.0107	-0.0091

CAPITAL COSTS (left label)

Table 6 Coefficient C_1, the Capacity and Costs Portion of the Linear Coefficient Resulting from Quadratic Estimation

INSTALLED CAPACITY IN 1997

	COAL	GAS	OIL	LWR	INT.COM.
NUCLEAR	1.5567	0.0000	-0.0001	-1.7952	2.4162
COAL	-2.6225	0.0000	29.1872	-1.0615	1.6244
OIL	-0.0855	0.0001	1.3163	-0.0781	0.3365
INT.COM.	0.0006	0.0001	0.0001	0.0064	-0.0236

CAPITAL COSTS (left label)

Table 7 Coefficient C_2, the Capacity and Costs Portion of the Quadratic Coefficient Resulting from Quadratic Estimation

One objection which has been stated about both Criterion Sensitivity and Describing Functions is that it is difficult to characterize the input uncertainties (i.e. define the set Ω_Δ or probability density $p(\Delta\alpha)$). It is true that such characterization is often difficult. However ignoring the input uncertainties just because they are difficult to characterize does not make those uncertainties go away. One advantage to the approaches discussed in this paper is that they provide explicit vehicles for dealing with input uncertainties, to the extent they are known, and in a systematic fashion.

The characterization of the input probability density enters the describing function approach in two ways. The probability density provides the "weights" for the weighted least square fits. If the resulting describing function is then used only for "plotting" and other such studies, an exact characterization of the input probability density is not essential. However if the mean and mean square of the describing functions are to be computed and interpreted, more care in specifying the input probability density is required. This might lead to the consideration of a series of studies where one probability density is used to get the weights for computing the describing function itself while parametric studies of various probability densities are done to see how the mean and mean square are affected.

The types of sensitivity analyses discussed here require the mathematical model (computer program) to be run many times. Some mathematical models of interest are so large and time-consuming that the costs involved in a requisite number of runs are prohibitive. There are two approaches in such cases. One involves decomposing and studying the model in great detail and then doing very limited sensitivity analysis on only those few input parameters that seem to be critical. This can yield meaningful results but it puts a tremendous burden upon the preparatory analysis, analysis which would require both an extensive a priori experience in the particular type of model and a comprehensive understanding of the model itself. It is very possible that this will not be a completely satisfactory undertaking. The second approach is to refuse to assess models that are very large and costly under the philosophy that no model can be assessed if it cannot be run often enough to be understood.

6. ACKNLOWEDGMENTS

Mr. McKay (Los Alamos): I have two questions. It appeared that you were looking at one at a time perturbations, which means we are looking effectively in directions parallel to coordinate axis, is that correct?

Dr. Schweppe: The numerical results that we had were one at a time, the periods very definitely not one at a time. It falls apart if it is not one at a time.

Mr. McKay: So that in your fit, you didn't really worry about cross-produ terms and in fact you probably did not include them in your response surface polynomial?

Dr. Schweppe: What we did actually on this project is, we had a limited amount of data, when we finally thought of doing all of this stuff, and began t understand what it was all about and that stage was about over, we only had a one at a time perturbation at one block of matrices; we only had that in the computer. We were not able to fit cross terms.

The quotation that I was giving on how much computer time it would cost, that w our best guess, involving many, many more runs that would include cross terms.

You are right, the results we actually computed did not have the cross terms in

Mr. McKay: When you used your surrogate model, this polynomial, how did y study it? You said "pictures" and what not, but did you look just at plots and curves, or did you look actually at coefficients?

Dr. Schweppe: We did not look at plots and curves, that is something I ju added as something we should have done. It is always more probably what one sh have done, not what one did. What we actually looked at is outputs of the mean and covariance of the output. You can compute from those coefficients the mean and the covariance of the output. So we had a covariance of the output. We ha a 15 by 15 covariance matrix of the uncertainty in the output. That is the onl thing we really looked at. Also the mean vector. That is the only thing we re looked at and to be honest with you towards the end of it we found more and mor problems with what we were doing. We finally discovered the right way to do it and the only thing we ever looked at that was really precise was runs where we took some of the earlier runs which we didn't like and fixed them up by hand. it was valid, but we never looked at the night's computer printout that had all the correct covariance matrices in it.

I am confident now that the technique does work but it was done by hand.

Mr. McKay: Did you just happen to look at the mean vector that you calcul from your fit and compare it to the mean of the data and the covariance matrix of your output variables to that that you derived from your fits?

Dr. Schweppe: For this particular REM case, the mean of the deviation was very small compared to the covariances, i.e. the mean was small. It is hard to put a value judgment on what is small, but it looked small.

7. REFERENCES

1. Joskow P.L. and M.L. Baughman, "The Future of the U.S. Nuclear Energy Industry," *Bell Journal of Economics*, Vol.7, No. 1, 1976.

2. MIT Model Assessment Group, "Independent Assessment of Energy Policy Models: Two Case Studies," MIT Energy Lab, Report MIT-EL 78-011, May 1978.

3. Cecilia Sau Yen Wong, "New Approach in Parameter Sensitivity for Model Assessment," MIT Energy Laboratory Working Paper, June 21, 1978.

A NEW APPROACH TO ANALYZE INFORMATION CONTAINED IN A MODEL

Harvey J. Greenberg
Office of Analysis Oversight and Access
Energy Information Administration

INTRODUCTION

The purpose of this paper is to summarize the development
of a new approach to address the general question: What
information is contained in a model? For example, the
equation, $E=mc^2$, is a model that relates to two variables,
energy (E) and mass (m), with a numerical constant (c^2).

The Energy Information Administration (EIA) is required
to provide not only numerical data, but relations among
data; not only historical measurements, but forecast
estimates; not only basic projections, but impacts of
proposed policies. Since energy information is complex,
analysis is imperfect, and decisions are difficult,
instructive use of energy information depends upon the
accuracy, reliability, and credibility by which the
information is recorded and interpreted.

Furthermore, since the scope of energy analysis affects
every person, industry, and environment, it is vital to
apply engineering and economic skills not only artfully,
but scientifically. The new approach proposes to account
for relational and numerical information with a unified
structure to record and analyze the information contained
in a model. Questions of information contents may pertain
not only to the explicit data that was recorded, but to
implied relations. For example, suppose a model relates
three processes: production, transportation, and consumption
of oil. Their amounts may be related, for example, to asso-
ciated prices at points of supply and demand. Figure 1
illustrates such a structure, where the constants, 1 and
-1, and the parameters, C_1, C_2, C_3, U_1, U_2, and U_3, comprise
the numerical data.[1]/

FIGURE 1

A PHYSICAL FLOWS MODEL

	Production	Transportation	Consumption
Supply	1	-1	
Demand		1	-1
Cost	C_1	C_2	C_3
Capacity	U_1	U_2	U_3

---

[1]/ One may think of the Physical Flows Model as a linear
program. The Supply and Demand rows then represent
"material balances," and the columns represent three
activities. The Cost and Capacity rows contain objective
and bound values, respectively.

The goal of the new approach is to be able to answer
questions pertaining to a model's implicit, as well as
explicit, information contents, for three forms of
analysis: validation, verification, and assessment.

A validation exercise may be concerned with comparing the
accuracy of the model's information contents with evidence
obtained from other sources, such as judgments from experts
or indications provided by historical trends. Verifica-
tion, on the other hand, deals with whether the model's
information contents agree with the documentation. Assess-
ment may be relative to other models that are designed to
represent the same numerical information but with different
relations. All three forms of analysis--validation, verifi-
cation, and assessment--require answers to questions pertaining
not only to the explicit information, but to the model's
implied relations--that is, the implicit information contents.

The new approach, which is described in a series of
technical memoranda (see references), proposes a unified
structure in two dimensions: the modeling framework and
the form of analysis. To indicate the extent of the unifi-
cation, the next section outlines the scope of the proposed
approach. Then, an overview of the constructs that com-
prise the new approach is presented. Focus is on three
related questions: How are relations defined?; How are
they determined?; and, How are they measured?

The conceptual approach, however, is only one of the
prerequisites for success. A second issue is whether the
proposal possesses sufficient rigor that it can be auto-
mated--that is, "Is it feasible to implement the approach?"
We are especially interested in large, complex models,
where the information is not readily apparent.

The concluding section summarizes the proposed approach
and its implementation. The central conclusion is that a
variety of modeling frameworks, including most used by
EIA, can be unified into a new form that organizes the
information into a useful structure. By applying cur-
rent computer technology, a system capable of answering
questions, retrieving information to validate, verify, and
assess a model during its development, application, and
audit is feasible to implement.

518

SCOPE

The new approach has two dimensions: the modeling
framework and the form of information analysis. Cur-
rently, there is no taxonomy for model structures;
nevertheless, different modeling techniques generally use
different accounting principles. A linear programming
model, for example, is oriented towards deterministic
representation of "activities" which must satisfy "con-
straints" as they comprise a "process." An econometric
model, however, is oriented towards "exogenous" or
"explanatory" variables to statistically estimate
"endogenous" or "dependent" variables. The proposal
unifies these two apparently opposite forms of repre-
sentation into one accounting structure: a "matricial
form."

The anatomy of a matricial form is comprised of constructs
that embody both relational and numerical information.
First, there is a set of variables that are divided into
two parts: rows and columns. Generally, a matricial
form has a specified number of variables (n), of which
a specified number (m) must be rows. We refer to their
difference (n-m) as the "degrees of freedom." Each of
the possible assignments of variables to be in the row,
vs. column, set constitutes a configuration of the
matricial form. The reason for considering different
configurations is to examine implied relations.

The first division represents relations between the two
sets of variables: rows and columns. For example, in
a linear programming model, the original matrix configu-
ration uses columns to represent activities and rows
to represent either constraints, objectives, or accumu-
lations for report-writing. A basis, such as at optimality,
corresponds to one of the configurations. By contrast, an
econometric model represents the explanatory variables as
columns and the dependent variables as rows. An alterna-
tive configuration describes implied relations, for example,
between two dependent variables.

The second division pertains to the meaning of elemental
values--that is, the location, sign, and magnitude of a
value at a row-column intersection. (In some cases, only
the locations of nonzeroes are known; in other cases,
only their locations and signs are known.) The location
of an elemental value generally relates the associated
row and column variables. However, some rows represent

variable-specific (or unitary) information--for example, a bound on the capacity of an activity or a range of fixed values for an explanatory variable. Furthermore, some columns contain information associated with the row variables--for example, nonzero entries in a system of equations. In general, the information represented by an elemental value may be unitary, or it may represent an interaction between two variables. The second division, therefore, defines two parts: body and rim. The body "embodies" relational information between row and column variables, and the rim contains unitary elemental values.

The body of a matricial form is a matrix. A question addressed by the proposed approach is: If the body contains only the locations and signs of elemental values, but not numerical data, is it still possible to determine implied relations? This question belongs to a class of problems, called "qualitative determinancy," which was posed by Paul Samuelson [7], and is analyzed with the new approach in reference [1].

A second question of interest is illustrated as follows. Suppose a model represents regional production, conversion, and consumption of petroleum products, as well as inter-regional transportation by pipeline or tanker. The model may have thousands of variables and may use many different databases, thus making its information contents impos-sible to comprehend, and perhaps prone to error, without some automated aid. Validation and verification exercises must examine the flow relations. For example, a query may be: Does the model account for flow of gasoline from Texas to New York? A question of causality is: Would an increases in Texas' refining capacity affect New York's gasoline supply? These two questions illustrate the need to organize the modeling framework into a structure cap-able of answering queries about the model's relations.

In summary, one measure of scope is the extent to which the model's information can be revealed for direct reporting A second dimension is the extent to which a "diagnostic aid" can be developed for certain applications. For example, a model may produce a fallacious result because it contains a faulty element, such as incorrectly entered data. The analyst must trace the result to its cause, often under severe time pressure. The proposed approach, once implemented, offers aid to the diagnostic analysis by automating the determination of causality.

UNIFYING PRINCIPLES

The purpose of this section is to summarize some of the developments obtained thus far.

First, a cardinal measure of economic correlation has been proposed and studied [1]. It is defined, relative to the choice of row (vs. column) variables, to be the inner product of the associated column vectors. The sign of this correlation determines an ordinal relation: two column variables are complements, substitutes, or independent, according to whether their economic correlation is negative, positive, or zero, respectively. Several classes of models were examined to test how well this measure captures intended relationships.

For example, Figure 1 illustrates a matricial form that represents physical flows from supply to demand. The columns are comprised of three classes: production, transportation, and consumption. Since the production and transportation columns intersect a common supply row with opposite signs, their economic correlation is negative, so they are complements. This means an increase in a region's production must be accompanied by an increase in outbound transportation. Furthermore, production and consumption columns are independent, relative to the choice of row variables shown, because they do not intersect a common row; however, there is another choice of row variables shown in Figure 2, where Production and Consumption appear as complements. This raises two related questions:

1. Can the proposed measure of economic correlation be extended to measure relationships for many configurations without explicity computing each one?

2. Are there interesting classes of models for which the sign of the economic correlation does not change over all configurations?

The answer to both questions is yes, and reference [1] develops the associated concepts and specific results.

Another form of implied information pertains to a problem posed by Koopmans and Bausch [6, p. 138]. The problem is to determine an "embedded hierarchy" that traces causation. For example, Figure 3 illustrates hierarchies of the three column variables for two configurations of the physical flows model. Extensions and solutions of this form of inferential analysis is given in reference [2]; algorithms to solve the associated search problem are described in [3,4].

FIGURE 2

RECONFIGURATION OF PHYSICAL FLOWS MODEL[a/]

	Production	Supply	Consumption
Transportation	–	–	
Demand	+	–	–

[a/] Only the body is shown with only the signs of the elemental values.

FIGURE 3

HIERARCHICAL RELATIONS FOR PHYSICAL
FLOWS EXAMPLE
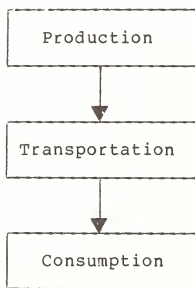
(a)

Original Form
(Figure 1)
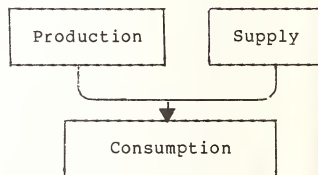
(b)

Reconfiguration
(Figure 2)

522

In summary, the matricial form accounting structure provides a unified modeling framework and enables the development of principles of diagnostic analysis. Particular problems, such as those posed by Paul Samuelson and Tjalling Koopmans, are readily solved by applying basic elements of graph theory and the algorithmic technology of computer science.

CONCLUSION

Two new constructs have been developed:

1. An accounting scheme, called a "matricial form"; and,

2. A process, called "diagnostic analysis."

The first provides a unified modeling framework; the second organizes information retrieval problems. With current computer technology, the approach may be implemented to comprise a system to validate, verify, and assess models.

DISCUSSION

Mr. McKay: Could you comment on applications, or lack of applications you see as the matrix becomes more and more dense?

Dr. Greenberg: I don't know of any problems that are on the order of 2,000 x 10,000 that are dense. I don't think that you could fit it in core with the current capabilities. I am interested in large-scale models. If it is 23 x 23, I figure you could probably get more information by drawing it on a piece of paper, and you wouldn't need the sophisticated software.

The only reason for having software for doing this is that it is too much information to fathom by eyeballing; so I am presuming that the whole thing is going to be applied to large systems. I actually don't think it is possible to have a large system and have it be dense. I just don't think it exists. I try not to worry about that which I don't think exists.

Mr. Graves (Resource Planning Associates): Do you have a way as yet of determining what bases you need to look at for a particular pair of variables?

Dr. Greenberg: No. So far, the basis part is working. I have looked at PIES bases, and there are interesting stories that I have, but I don't think there is time to discuss the results here. I think after the speakers are finished, we could go out in the hall, and I could tell you about some of the things I have looked at and made some interesting observations by applying it to PIES.

REFERENCES

Technical Memoranda under preparation by Harvey J. Greenberg:

[1]   Diagnostic Analysis of Matricial Forms: Measuring Complementarity and Qualitative Determancy;

[2]   Diagnostic Analysis of Matricial Forms: Hierarchial Relations;

[3]   Information Structures for Matricial Forms;

[4]   Traceback Algorithms for Infeasible Matricial Forms; and,

[5]   Primer and User Guide for a Diagnostic Online Evaluation System (DOES).

Other references:

[6]   Koopmans, T.C., and A.F. Bausch, "Selected Topics in Economics Involving Mathematical Reasoning," SIAM Review, Vol. 1, No. 2, July 1959, pp. 79-148.

[7]   Samuelson, P.A., The Foundations of Economic Analysis, Harvard University Press, Cambridge, Massachusetts, 1955, pp. 23-28.

VALIDATING THE HIRST RESIDENTIAL ENERGY USE /
MID-RANGE ENERGY FORECASTING SYSTEM INTERFACE

Frank Hopkins
Energy Information Administration
Department of Energy
Washington, D. C.

Lewis Rubin
Electric Power Research Institute
Palo Alto, California

I. INTRODUCTION

There are multiple goals in a detailed model assessment: data validation,
software validation, documentation, etc. The most important may be docu-
mentation of the assessment process since this permits outside verifica-
tion of the assessment results. The documentation should be highly
detailed, but it must also maintain an overview perspective of its role
as an identifier of the fundamental underlying relations which give the
model its behavioral character.

In another vein, documentation is a very valuable tool for use in inte-
grating a particular model into an existing modeling network. It is
precisely this type of information, that the documentation aspect of
a model assessment produces, which is of critical value to the inte-
gration process.

An assessment, for example, would uncover the following types of infor-
mation about a given model:

   o Dimensionalities, such as region and sector structures;

   o Time frames;

   o Data sources and definitions;

   o An identification of exogenous and endogenous variables; and,

   o Some notion of the essential skeletal structure of the model,
      an understanding of which is critical to the form reduction
      process so often necessary in model integration.

We would like to explore these notions by means of an example of the
model integration process, which we recently completed to incorporate
the Hirst Residential Model into the DOE Mid-Range Energy Forecasting
System (MEFS).[1] This effort was started without benefit of a compre-
hensive assessment document as a reference tool; thus our contention
that assessment work is helpful, if not critical to a model integration
process, is borne of real and painful experience.

In the next section we would like to discuss in more detail the definition
of model integration and present some basic concepts on the integration
process. The third section of the paper will address why model integration
is of interest and importance in the context of the policy analysis process.

---

[1] The Project Independence Evaluation System (PIES) was the predecessor
     to the DOE Mid-Range Energy Forecasting System. The component models
     of PIES have been substantially modified since the initial use of PIES
     in 1974, so that the original name does not adequately describe the
     current system.

And finally, the fourth section of the paper will attempt to explain how assessment can help by illustrating specific problems and solutions from the Hirst model integration effort. The conclusion will summarize the major results of the study.

II. WHAT IS MODEL INTEGRATION?

Model integration describes a process by which two or more conceptually complementary models are linked in order to produce a more comprehensive analysis then could be accomplished by the individual models. Typically, models which are complementary describe different but related subsets of a physical or social system.

A macroeconomist, for example, might be interested in linking a national income model to an interindustry model of the economy. In so doing, he might explore two areas more fully: (1) the interrelationships between consumption and investment, and (2) changes in factor demands by industry. This analysis would be infeasible using each model individually. Once they are linked, however, they form a powerful tool for understanding integrated analyses.

In energy analysis, the operative system is typically the energy market network. Complementary models include those which describe primary supply of coal, oil, gas, transformation activities, and demands. To analyze electric utility capacity and generation decisions in the absence of reasonable electricity demand forecast, for example, is not typically useful. Also further attempts to forecast electricity demand in the absence of a reasonable description of natural gas demand is not meaningful.

Conversely, models which are not complementary are competitive. Competitive models describe the very same phenomena (i.e., the same piece of the network) using different techniques or different degrees of detail. The two models being discussed in this paper are obviously competitive, which is why the exercise being described involved replacing one with the other at a particular node in the DOE energy modeling network.

Model compatibility is a necessary condition for model integration. If two models are not compatible, their characterizations of their subset of a general system are not fusible in relation to each other. Furthermore, noncompatible models may not have a feasible mechanism for passing information between each model and thus reconciling feedback effects becomes impossible. As such, they can not function in tandem and are unlinkable.

Thus, model integration is also a process by which models are modified so they will be compatible with one another. There are two major ways in which models must be altered in order to make them compatible:

o Reconcilication in terms of their data definitions and accounting structures; and,

o Implementation of common solution algorithms.

Data definition refers to the data sources used in a model and the procedures for defining internal variables. Industrial gas demand, for example, includes or does not include refinery demand. It can also include or not include feedstock demand. It can include or not include lease and plant fuel.

These concepts become particularly important in the context of a model integration exercise because of the accounting structures imposed on modeling systems. Each model within the system is no longer free to independently define the nature and content of its internal variables, because the system demands material balance integrity. What goes in must come out and nothing can be counted twice. If these rules are violated, the modeling system can produce disastrous distortions.

Consider, for example, an energy system in which all supply models utilize the Bureau of Mines (BOM) data definitions, but the industrial demand model observes Census of Manufacturers definitions. Because the primary data collection coverage of the two agencies is different, the Census of Manufacturers is typically between 5 and 30 percent lower than BOM. Thus, industrial demands will end up being significantly understated. Solution of the model will generate a price/quantity equilibrium that may yield a zero level of imports in 1990 because of the understated demands. This may be a tempting way to solve the energy crisis, but it is not science, and it certainly is not policy analysis.

The algorithms used to solve systems of linked models are important because they specify the ways in which information is passed among the models. Without the information exchange, of course, feedbacks cannot be reconciled and the integration process is not complete.

Typically, the algorithms used to link models are complex because the individual models themselves are of diverse types. In the DOE's MEFS model, for example, engineering, econometric, and optimization-type models are linked within a linear programming/simultaneous equation framework to solve the model. In the generalized equilibrium modeling framework, of which the Gulf-SRI model is the best known example, a network algorithm is employed in which prices and quantities are the only kinds of information passed between models. Individual nodes within the network can be represented by optimization, simulation, or detailed structural models as long as prices and quantities can be extracted from their results.

527

Diversity of inputs utilized in a modeling system demands an algorithmic discipline be imposed on the individual models. Algorithmic structures demand conformity similiar to those required in accounting structures. In essence, the way in which a model is represented to the system is controlled in terms of variables, time frames, and equation specifications. This may not, however, be the way the model is represented in free-standing mode, in such a case, a certain type of alteration known as "form-reduction" or "cartooning" 2/ must be undertaken. This process will be discussed in detail in the next section. The "form-reduction" process is the basic way in which models are made algorithmically compatible for integration.

In summary, the question of what model integration is has been answered. In addition, several general principles of the process have been delineated. But the question of why model integration is important is still unanswered. In particular, how is integration important to the policy analysis process, and why is this technique preferable to the separate construction of large multimodal models tailored to each policy question. These issues will be addressed in the next section.

III. THE ROLE OF MODEL INTEGRATION IN THE POLICY ANALYSIS PROCESS

Model integration is distinct from model building. However, the two activities have identical goals with respect to the policy analysis process. But model integration adds an extra dimension to the modeling facility by allowing greater speed and flexibility in the development of comprehensive analytical tools.

Essentially, model integration permits the construction of integrated analysis tools from prefabricated pieces. This creates broad possibilities for specialization of labor in model building, quick and easy redirection of the emphasis of model design and function, and timely improvement and updating of models. These facilities are important to the policy analysis process because of the interrelationship among three factors: the importance of detail, the importance of comprehensiveness, and the importance of producing timely analysis and monitoring an operationally manageable system.

Detail - Detail is of critical importance in the evaluation of policy decisions precisely because most policy decisions concern themselves with disaggregated areas within the social or economic system. In energy policy, this has become increasingly true as the complexity of

2/ This term is attributed to David Nissen of the Chase Manhattan Bank.

the issues has been recognized over time. There is no doubt that the days of blanket BTU taxes and gross import quotas are gone, essentially because such policies were abandoned as unworkable in a complex energy/economy system.

Detail analysis is required by the following types of questions addressed to the Energy Information Administration (EIA) of DOE:

o How much energy savings should be expected in 1990 from a 20 percent tax credit on home storm windows;

o How much more (or less) natural gas demand and prices should be expected in 1990 if gas is deregulated and is priced incrementally to residential customers;

o How effectively can industrial customers be switched out of oil and gas and into coal by 1990 through the implementation of a 30 percent tax credit on new coal equipment; and,

o Which industries will be affected and how much will such changes cost?

In another, but similar vein, the Electric Power Research Institute must answer questions like:

o What sorts of new technologies are going to be most cost-effective for electric utilities in the year 2000 and beyond;

o How effective will load management and time-of-day rate policies be in reducing the capacity requirements of utilities in the future; and,

o What is the role of energy storage expected to be, and will it be competitive with other types of technology as a method for managing the peak-load problem?

Clearly, questions such as these demand an enormous capacity for detail in the analytical process. As a result, energy modeling has continually evolved in the direction of bigger, more detailed models in order to accommodate its clients.

Comprehensiveness - The need for comprehensiveness has become increasingly clear in the process of energy policy analysis. Interrelationships among submarkets of the energy/economy system have manifested themselves time and again to policymakers and analysts, and the need to account for secondary effects of most policy proposals is well established.

In the broadest sense, feedback effects between the energy sector and the rest of the economy have been widely investigated since the first oil embargo of 1973. It was recognized that the essential problem posed by the energy crisis was much more complex than merely how to save energy. If energy consumption growth was cut, would GNP growth also be reduced? To what extent were these two growth rates coupled? What kinds of policies might be designed to uncouple them?

Comprehensiveness is also recognized to be important in other more specific areas within energy/economy markets. For example, the guzzler tax proposed as part of the original National Energy Plan was perceived through analysis to have two effects. Primarily, it was designed to entice motorists to switch to more efficient cars by taxing inefficient auto purchases and offering rebates for purchases of efficient ones. But a secondary consideration, about which the Treasury Department was very concerned during the course of the actual legislative design, was whether or not the tax revenues of this policy would wholly finance its rebate payments.

Still other questions abound. Can electric utility capacity expansion and dispatching decisions be sensibly analyzed without some clear idea of their interrelationships with electricity demand? Further, what is the nature of the interrelationship between electricity demand and natural gas demand? Can a policy affecting either be sensibly analyzed without some notion of how that policy affects the other?

Comprehensive modeling systems are particularly critical when attempting to analyze composite effects. If a coal conversion program saves two quads of industrial gas consumption in 1990, and an incremental gas pricing scheme saves two quads of residential gas consumption, how much total gas is saved by implementing both policies? Extensive experience with the MEFS model has demonstrated that the answer is rarely four quads, because sectoral interactions do not allow the savings to be fully additive. Such "cancelling" effects can easily be missed without a comprehensive modeling tool, and policy effects can be misinterpreted as a result.

Timeliness - Finally, it should be recognized that modeling in support of the policy analysis process, must be easily redirected to address topical issues; and it must produce easily interpreted results in a timely manner. The policy analysis activity typically demands prompt answers to its questions, and will only utilize tools which provide these answers quickly and easily. This is particularly true as regards the history of analysis in support of the Federal Government's energy policy process. During the planning and design of the original National Energy Plan in the spring of 1977, new energy policies concerning all aspects of supply and demand were being proposed and tested virtually on a daily basis. As the MEFS model was the chief modeling tool used in

that process, both its modeling structure and its software/data base system were put to stringent adaptability tests throughout the period. The extent of MEFS's flexibility at this time was one of the major factors determining the extent to which it contributed to the design of the National Energy Plan.

Thus, energy modeling must be responsive to its clients. It cannot allow itself to grow ponderous and sluggish or it will no longer be of use. And it is precisely the model integration mechanism which provides modelers with the added flexibility and redirective capability needed to be responsive in such an environment.

Model integration techniques allow a reconciliation between the need for complex, highly detailed, yet fully comprehensive models and the need for flexibility and speed of response. Rather than actually build- ing big, complex models for specialty purposes, analysts need only develop generalized structures (such as the SRI-Gulf energy network structure) and the associated generalized computer software and data base management systems.

Concurrently, highly detailed representations of individual submarkets within the system can be developed by others (i.e., Hirst residential energy demand, Baughman-Joskow utility). Finally, the two activities can be pulled together through integration.

Further flexibility is achieved through the process of "cartooning" models. Theoretically, it is possible to "reduce out" one or more dimensions of any highly detailed model and capture its essential behavior with a highly simplified "cartoon" structure. Thus, in any given integrated analysis, modelers have the option of using either a detailed representation of a particular sub-market or its cartoon. If the analysis being done is specifically related to the sub-market in question, the detailed model should be used. If not, the cartoon is probably sufficient to track gross feedbacks.

All of the feeder models integrated into the MEFS system are "car- tooned" to some degree, and the specific procedures undertaken for the Hirst model effort will be discussed in the next section. It is sufficient at this point to note how powerful such a technique can be, and how it introduces closure on the model integration process.

It is the twin techniques of integration and cartooning, carefully and selectively applied, which allow the previously stated goals of detail, comprehensiveness, and manageability to be reconciled. These techniques allow the modeling tool to be completely modular, and further allow specific pieces to take on greater or lesser detail as needs arise. Thus, the model need never be any bigger than it has

to be for a specific purpose, and redirecting that purpose in the pro-
duction mode becomes virtually a pushbutton task.

Of course, no modeling system realizes such flexibility yet, either
in conceptual design or in software/data base design. But several,
including MEFS, are working toward it. The next section presents
detailed discussion of the results of a specific model integration
project and how a previously completed model assessment could have
assisted in this task.

IV. ASSESSMENT REQUIREMENTS FOR MODEL INTEGRATION

This section describes a generalized process for assessing and modifying
models that are integrated into a larger modeling framework. The general
attributes of the model that must be assessed can be divided into five
categories: time period, dimensionalities, data definition, reduced form
characterization of the structural model, and endogenous and exogenous
variables. The modification of the models is always in the direction of
changing the small model to make it conform to the larger modeling system.
Each of these attributes will be discussed in both general terms and
using illustrations from the integration of the Hirst Structural
Residential Energy Use Model into the MEFS.

Time Frame

The general integration process must reconcile differences in time
periods used in each modeling system. These differences can be cate-
gorized into two groups: model structure and time dimension. The time
frame in model structure refers to the time phase of forecasted events
in a model simulation. Models can either be dynamic or static. Dynamic
models can be temporally independent or interdependent via lagged endog-
enous variables. The time dimension refers to the correspondence between
dates or periods between two models. These factors include initialization
dates, forecast period (annual, decennial, etc.) and forecast duration.

The interface requirements of the MEFS require static supply and demand
curves at 5-year intervals. The MEFS is solved for 1985, 1990, and 1995.
Both RDFOR and the Hirst model are annual recursive dynamic models,
both capable of generating demand forecasts to 2000. The Hirst model is
initialized using 1970 data, while RDFOR within MEFS, is initialized with
1977 data. Thus, the general preparation of the demand curves consists
of representing the dynamic demand function as a static generalized price

elasticity estimate and an initial set of quantities, specified for a scenario run of the demand model.[3] Since both models have been extensively documented elsewhere, Hirst and Carney [1] and MacRae [2], they will not be described in detail here.

Preprocessing is required so that the MEFS will not be required to solve RDFOR endogenously, which would be prohibitive in both computer storage and CPU time. The dynamic RDFOR model is used as a preprocessor to the integrating model to create a set of snapshot demand curves of the form:

$$(1) \quad Q_{i, t} = K_{i, t} * P_{i, t}^{Ei} * P_{j, t}^{Ej}$$

where i, j = fuel $i \neq j$

t = time

P = energy price

E = elasticity parameters

K = constant term embodying initial conditions

In (1) fuel consumption in year t is a function of own and cross price elasticities and initial conditions. The initial conditions reflect macroeconomic assumptions, weather and the dynamic behavior of energy prices that are used as scenario inputs into RDFOR. Thus, a solution of MEFS implies an absense of a feedback between the energy and economic sectors, normal weather and a prespecified set of dynamic prices. The constant elasticity specification of (1) is the only functional form that can currently be used in the MEFS. While it is theoretically possible to use other functional forms, the software for accepting non-constant elasticity demand curves has not been developed.

The translation of the dynamic RDFOR model into the static model (1) is accomplished by running RDFOR for a base price path and a pertubated price path as illustrated in Figure 1. The price differential is used to calculate static elasticities between the price paths for the year

---

[3] Lawrence Lau and Dennis Fromholzer and conducting a similar study to integrate the Hirst model into the Fossil II model developed by Roger Naill for the Office of Policy and Evaluation at DOE. The Fossil II model requires dynamic reduced form demand curves in contrast to the static or snapshot curves required by MEFS.

that is used in (1). The level of demand on the base price path is used to scale the constant term in the equation.

Static elasticities can be derived for each year in the simulation period by examining quantities derived from the base and perturbed price paths. Because of the lagged dependent variable specification of the demand model, the elasticity increases over time. This result is consistent with the ability of consumers to increase their response to price changes the longer the price change has been in effect.

Finally, it should be noted that while MEFS projects a set of equilibrium prices in a future year, it has no capability for determining the path which prices must take to reach that equilibrium. The choice of a price path is exogenous to initiating the model. The demand levels in future years will be dependent upon both the level and the historical growth rate of prices.

Hirst – MEFS Interface Requirements – The previous discussion concerned itself with issues related to integrating the reduced-form demand model into a functional form convenient to the MEFS integrating model algorithm Integrating the new Hirst residential model is a more complex task, because of differences in functional forms and product definitions.

RDFOR does not lose a great deal of accuracy in the process of being distilled into the simplified structure because constant elasticity static demand curves are a natural outgrowth of the original estimation process. However, the more detailed SEED models may prove to have quite different integrability characteristics. The use of the constant elasticity functional form for the Hirst model is questionable. Unlike the old, reduced form residential sector, the Hirst model system is not exclusively log-linear in its specifications. The dynamic relationships between prices and quantities cannot be analytically derived; yet it is necessary to approximate them by a static constant elasticity function to satisfy the MEFS integrating algorithm. Two approaches were used to generate the approximation. The first utilized multiple runs of the Hirst model to generate pseudo data to be used to estimate demand curves using regression analysis. The second used an elasticity approximation approach similar to the RDFOR-MEFS interface system.

Model Dimensions

The dimensions of the list of forecasted variables of both models must be assessed before integration can proceed. The regional and sectorial structures are the major dimensions that should be examined in model integration. The Hirst model generates demand curves for the 10 Federal regions, while the RDFOR forecasts are for the 10 Federal regions estimated from a 50 region state level model.

FIGURE 1

AN ILLUSTRATION OF THE RAMP PRICE PERTURBATION FOR
ELASTICITY CALCULATIONS
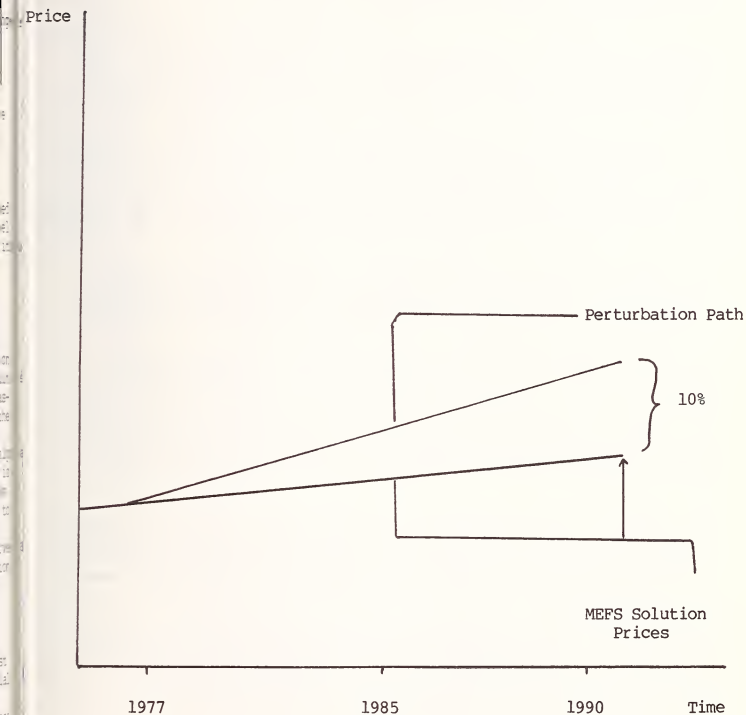
The RDFOR model has only one sector for total residential energy use, disaggregated by four major fuel types: electricity, natural gas, distillate oil, and liquid gases. Coal used for home heating is classified in the minor fuels category in RDFOR. The Hirst model has three major fuels and an other category: electricity, natural gas, and distillate oil. Liquid gases are included in the other category. The Hirst model has been expanded to include liquid gases separately to achieve sectorial consistency with MEFS.

The Hirst model also provides more end-use detail than RDFOR. Energy demand is forecast for single-family homes, multi-family homes, and mobile homes in Hirst, while RDFOR only forecasts total residential energy demand.

The regional version of the Hirst model constructed at ORNL used the same national energy use elasticities for all regions. Initial simulations of this version revealed that while the national totals were satisfactory, the regional detail was questionable. The RDFOR regional elasticities were used to replace the national elasticities in an attempt to improve the regional representation of the model.

The elasticities of the Hirst model are disaggregated from a total elasticity measure into three components: new equipment ownership (market share) elasticities, equipment usage elasticities, and equipment efficiency elasticities.

The new equipment market shares elasticities are detailed own and cross-price and income elasticities which relate new equipment capital choices in the residential sector to changes in fuel prices. For example, if the price of electricity increases, the share of new electrical equipment for a given end-use will decrease, while the share of new equipment using alternative fuels for the same end-use will increase.

The usage elasticity is composed of own-price and income elasticities which report how the usage of various residential capital equipment is affected by the price of fuel and by levels of income. For example, if the price of gas increases, owners of gas space heaters will reduce their thermostats.

Equipment efficiency elasticities are technical parameters which relate improvements in the efficiency of residential capital to increases in fuel cost. They remain fixed across the regions of the country on the assumption that technology is not regionally dependent.

These three component elasticities (market share, usage, and equipment efficiency) can be chosen in such a way as to be compatible with overall econometrically estimated elasticities using the Elasticity Estimator program developed at Oak Ridge. RDFOR-generated regional elasticities were entered as control totals to generate the three component elasticities that are consistent with the RDFOR regional control elasticities, to replace the original nationally based elasticities.

Reconciling Model Product Definitions

The first problem that was resolved in integrating the Hirst residential model into the overall MEFS structure was that of product definition. In the RDFOR accounting framework, residential fuel consists of only those demands which are explicitly reported in the data sources as being residential. Natural gas and electricity are reported by blockrate categories, with one of the categories directly named residential. Distillate oil is classified as heating oil in the Bureau of Mines data, and it is further disaggregated into residential and commercial using proportions derived from the 1970 Census of Housing. The Hirst model, however, defines the residential sector products in the following ways:

- o An explicit adjustment in the Hirst model is made for natural gas and electricity for fuel used in gang-metered apartment buildings, which is reported as commercial use in RDFOR, but in reality is residential use. Thus, the Hirst fuel quantities are conceptually larger, including an amount which RDFOR assigns to the commercial sector. Double counting will occur if the proper adjustments are not completed.

- o The Hirst model includes an "other fuels" category which is primarily made up of liquid gas demand, but also includes coal and certain computational adjustments. These components must also be adjusted so they are equivalent to the RDFOR demand definitions.

These adjustments were completed through a reinitialization procedure. This procedure eliminates the necessity to reformulate the model using a reconstructed historical data base. Implicit in this approach is the assumption that there are no behavioral differences between the two definitions of the residential sector, that is, that the sum of detached home dwellers and gang-metered apartment dwellers behave identically to the set of single family residences. If the differences in fuel magnitudes are a small percentage of total demands (which is the case in the present exercise), this assumption should not result in a serious distortion in the projections. The two procedures used to recalibrate the Hirst model initial conditions to RDFOR sectoral definitions follow.

First, EU, which represents energy used by a typical unit (appliance) in the Hirst model base year (1970) is specified as part of the initial conditions. These values were adjusted for market penetration, reconciled to the housing stock data, and aggregated by end-use to yield total quantities of each fuel consumed in 1970, consistent with the RDFOR totals by the procedure discussed in the next step.

Second, when the model is simulated, the corrected 1970 data generates a projection of total fuel consumed for 1977 that differs quantitatively from the 1977 RDFOR historical data.

A set of factors is calculated to reconcile the Hirst simulated 1977 data to the RDFOR historical data on a fuel-by-buel basis. Table 1 presents the 1977 fuel use values as defined by Hirst and RDFOR, as well as the set of factors used in the recalibration of the EU factors.

Model Reduction

The integration of the Hirst model was only computationally feasible if the large structure could be reduced to the constant elasticity form (1). There were two procedures that were used in the reduction experiments: multiple pseudo data generation and single pair model perturbation. The results of the experiments demonstrated that estimating demand curves using psuedo data is too expensive for successful use in a production run of an integrated modeling system.

Psuedo Data – The pseudo data was generated by running the Hirst model by perturbating one fuel price to a base ramp price path, as illustrated in Figure 1, while holding the other prices constant. The regressions were run for each type by pooling the data across electricity, natural gas and distillate prices.

The initial experiment was run with perturbations from 50 percent below to 100 percent above the base price path at 2 percent intervals. To determine the robustness of the estimated elasticities, the sample was divided into three segments:

1. Prices 50 percent below to the base price path;

2. Prices from the base path to 50 percent above; and,

3. Prices 50 percent above to 100 percent above the base price path.

TABLE 1

NORMALIZATION FACTORS TO ADJUST HIRST DATA TO FFDS WEATHER ADJUSTED CONCEPTS
(Uses 8/24 Elasticities)

Region	OIL MEFS	OIL Hirst	OIL Factor	NATURAL GAS MEFS	NATURAL GAS Hirst	NATURAL GAS Factor
1	402.9	402.8	1.0003	147.0	157.3	.9344
2	616.5	594.6	1.0368	489.0	537.3	.9101
3	309.2	329.9	.9373	460.4	583.7	.7888
4	173.0	193.8	.8928	358.7	488.3	.7346
5	553.4	471.1	1.1747	1616.1	1902.1	.8496
6	68.7	49.3	1.3937	527.5	576.7	.9146
7	62.7	61.3	1.0220	400.0	529.8	.7551
8	27.1	33.5	.8075	205.2	272.7	.7524
9	34.1	10.9	3.1188	659.0	857.4	.7685
10	67.9	57.0	1.1930	68.7	101.0	.6809
U.S.	2315.5	2204.2	1.0505	4931.6	6006.3	.8211

Region	ELECTRICITY MEFS	ELECTRICITY Hirst	ELECTRICITY Factor	OTHER MEFS	OTHER Hirst	OTHER Factor
1	95.9	106.1	.904	18.2	14.9	1.2209
2	154.1	155.1	.9971	18.8	31.7	.5933
3	228.2	216.2	1.0618	24.6	72.1	.3415
4	513.0	527.1	.9674	114.7	144.5	.7940
5	417.0	405.7	1.0244	154.3	134.7	1.1453
6	278.2	280.7	.9685	103.4	120.5	.8585
7	122.3	126.1	.9505	99.0	85.1	1.1639
8	57.0	55.5	1.0256	37.2	34.2	1.0857
9	207.3	231.6	.8828	17.2	37.1	.4629
10	142.9	165.6	.8628	7.3	22.2	.3283
.S.	2215.9	2269.7	.9763	594.7	697.0	.8532

The R-squared values all exceed .99 for each regression in both the log and linear form. There were variations in the estimated elasticities in the different price ranges. While the own price elasticities increase over the higher price ranges for electricity and natural gas and decline for oil, their values are within 8 percent of the 50 percent to 100 percent perturbation range. The major variations occur in the cross price elasticity. The cross price elasticity estimates for oil demand have the largest variation.

The nonconstant price elasticity estimates generated over wide price ranges for the national model indicated that narrower price ranges should be used in estimating the regional demand curves. The regression analysis was performed on 31 pseudo data points over the 10 regions by perturbing each fuel price by 2 percent intervals, from 30 percent above the base price to 30 percent below the base price. This procedure required 930 runs of the Hirst model since it had to be perturbed over 10 regions, 3 fuels, and 31 prices.

Both the linear and double log regression results had high R-squared values, and statistically significant coefficients of the expected magnitude and signs. However an examination of the residuals of both functional forms showed that the error terms were autocorrelated. Thus, the demand curves could not be assumed to be constant elasticity. In the second experiment, arc elasticities were calculated between each of the 2 percent increments. Visual examination of the 31 arc elasticities demonstrated that the demand curve was not of constant elasticity.

Model Perturbation - The initial plan of using the multiple perturbations of the Hirst residential model in RDFOR to generate prices and quantities to estimate a demand curve of a constant elasticity form from pseudo data was abandoned for three reasons. First, simulations did not generate constant elasticity demand curves that were required by MEFS. Second, the perturbated Hirst runs require about 148 minutes of CPU time. This is an excessive amount of CPU time that could not be tolerated for a regular production run. Third, the regression equations generate an inaccurate representation of the arc elasticity around the base price particularly if the elasticity generated by a regression equation differs from the arc elasticity calculated along segments of the curve. Since the base price is the expected price at which equilibrium will be obtained, there is a low probability that equilibrium will occur at an appreciable distance from the base price.

The current procedure for generating a constant elasticity estimate is to use the RDFOR procedure illustrated in Figure 1. This technique reduced computer time so that flexibility in scheduling production runs was maintained. However, it must be recognized that the elasticity number that is to be passed to the MEFS loses reliability the further the equilibrium price is from the base price.

Exogenous and Endogenous Variables

The examination of the Hirst and RDFOR models, with respect to both
exogenous and endogenous generated information, was crucial to the
integration process. First, the exogenous inputs required to drive
the Hirst model were more extensive than those required for RDFOR.
Thus, the development of a system for generating exogenous forecasts
is a necessary condition for any model integration effort. Second,
as the Hirst model illustrates, models typically forecast more variables
than are required for the integration process. While these variables
may not be used in the integrated system, they serve a useful function
in assisting in validating the forecast results.

The reduced form RDFOR model has a less complex form than the struc-
tural Hirst model. Income, population, and energy prices are the major
exogenous variables, with income adjusted for population being the primary
deterrent of the level of demand. The Hirst model requires a signifi-
cantly larger set of exogenous variables: energy prices, energy equip-
ment prices, equipment efficiency and average life, interest rates,
housing stock and per capita income. In contrast to RDFOR, housing
stock and not income is the primary determinant of energy demand.

A housing model in the Hirst model system forecasts the demand for
occupied housing units as a function of exogenous income, population,
and household formation forecasts. It further forecasts increases in
average housing size as a function of income. Therefore, in integrating
the Hirst model into the general MEFS, consistency between the housing
forecasts and the overall macroeconomic forecasts obtained from DRI,
which drives all other sectors of the demand model, must be achieved.

As a prelude to developing a software system to assure this consistency,
a number of alternative housing forecasts were run to determine the
regional sensitivity of the model. An improved regional set of housing
forecasts, using the Hirst housing preprocessor, was developed from a
State-level version of the Hirst model. The Tetra-Tech group generated
regional population forecasts by reconciling them with current OBERS'
1977 based growth projections. In addition, they developed a method of
regionalizing headship rates (headship rate is defined as a number of
household heads per capita), rather than relying on the constant national
headship rate applied to all regions as in the original Hirst model.

A resident version of the housing preprocessor, complete with a direct
data linkage to the Hirst model input files, has been incorporated into
the demand model production stream to ensure macroeconomic consistency
with the other demand models. Thus, changes in macroeconomic scenarios
specified by Applied Analysis clients will be directly translated into
housing forecasts in the residential sector, as the macro forecast are
currently used directly and automatically in the other sectors.

In the future, a closer examination and improved specification of the entire housing model structure remains a fruitful area of model improvement.

- o As presently formulated, the model is exclusively demand-driven, and would be improved if supply factors subject to macroeconomic environments were incorporated in the model.

- o The development of a housing supply representation might facilitate the forecasting of fuel-specific housing types which is not feasible with the current housing model.

The Hirst model forecasts a large set of variables not required for the integration process--eight categories of appliance sales and stock, and energy use by fuel type for the three types of residential structures. While this information must be aggregated before it can be used in the MEFS, it provides a useful check on validating a forecast. The sales information could be used to provide a consistent interface with a macroeconomic model to analyze macroeconomic impacts of energy policy.

A series of test runs were made in order to empirically evaluate the performance of the Hirst residential model and the RDFOR system. Both models were run with identical income and energy price paths, initialized from 1977 historical data. Since the purpose of the simulation was model comparisons, conservation programs under current law were not included in the run of the Hirst model. Thus, the projections should not be considered meaningful EIA forecasts, but are only included for illustrative purposes. A comparison of national demand levels for both models for 1990 are presented in columns 1 and 2 of Table 2.

Two phenomena appear to summarize the differences in behavior. The total level of residential energy demand is .9 quads higher in Hirst than in RDFOR. The share of natural gas demand in Hirst is higher by 1.2 quads and accounts for most of this difference. Electricity and oil are slightly lower in Hirst, while the other fuel category is lower by .3 quad.

While the differences do not imply which model is more accurate, the higher levels of Hirst can be partially explained by the treatment of normal weather conditions, an exogenous variable, and of the moratorium on residential gas hookups. Appliance sales by fuel type are an endogenous variable within the Hirst model.

TABLE 2

HIRST ADJUSTMENT RUNS

1990 FORECASTED U.S. CONSUMPTION
(Trillion BTU's)

	Base RDFOR	Hirst	Weather Adjustment	With Gas Hookup Moratorium Only	Both Adjustments
Electricity	3952	3909	3857	4658	4617
Gas	4300	5463	5422	4304	4260
Oil	2371	2326	2267	2610	2546
Other	644	442	429	471	460
TOTAL	11267	12140	11975	12043	11883

Weather Adjustment – The Hirst model was initialized in 1970. The winter that year was colder than normal in all regions except DOE9, and was hotter than usual during the summer in all regions except DOE4 and DOE6. The forecasts in Table 2 were not adjusted to account for abnormal weather during the initialization year. To adjust the model for 30-year average weather, space heating and air conditioning were scaled in each region by a proportion of 30-year average weather to 1970 weather.

In the original runs of the Hirst model, consumption forecasts were benchmarked to 1977 actual levels of consumption. Since 1977 was a colder than normal winter, the weather adjustment to 30-year average resulted in a lower level of consumption than the model run bench-marked to 1977 actual levels of consumption, as illustrated in the third column of Table 2. The base case run of the Hirst model resulted in a level of consumption approximately .165 quad more than with 30-year average weather conditions imposed. The adjustment for 1977 weather is not large. However, if the initialization year has unusual weather, the difference can be significant.

Moratorium on Residential Gas Hookups – The market share equation for appliances assumes that consumers will base their appliance choice between electricity, natural gas, oil and propane upon income, equipment and fuel prices and appliance efficiency. The version of the model used in the initial simulation does not incorporate the moratorium on new residential gas hookups that have been occurring in all regions except DOE6 since 1972. RDFOR uses econometrically estimated equations using 1960-1975 data. Thus, a portion of the moratorium is incorporated in the forecasts. Hence, a true comparison of the Hirst model and RDFOR must include an adjustment in Hirst to reflect the moratorium.

A routine was added to the Hirst model code to prohibit new homes from using natural gas appliances in all regions except DOE6. Because of the way the switch was implemented, the relative growth rates of electricity and oil consumption were the same as in a model run without a moratorium. Thus, the switch from natural gas was mainly into electricity as illus-trated in the fourth column of Table 2. This simulation is reported without adjusting the model for normal weather conditions. In the mor-atorium case only, forecasted natural gas consumption was down 1.16 quads in 1990 for the U.S. using Hirst, and electricity and oil were up .7 quads and .3 quads, respectively. Therefore, total end-use consump-tion was down .1 quad.

Revised Simulations – The last column of Table 2 shows the result of running the model to incorporate both adjustments. These figures reflect modifications in the Hirst model to make it comparable to RDFOR, includ-ing definitional modifications, regional elasticity, weather and gas moratorium adjustments. Comparison of column five with column one, the

RDFOR base figures, shows that there is only a small difference of .6 quads. Projected natural gas consumption would be reduced slightly by .04 quads. Oil consumption would increase slightly by .18 quads. The major projected change would be an increase in electricity consumption by .67 quads using the fully modified Hirst model.

The combined adjustments in the modified Hirst decrease the original Hirst-RDFOR difference in 1990 natural gas consumption from 1.2 quads to -.04 quads. However, electricity consumption which had only been approximately equal between the two versions of the model is now .67 quads higher.

V. CONCLUSION

This paper has reviewed the importance of model assessment in integrating two models. The assessment process reveals critical factors that must be known before the integration process can be initiated. The assessment process also provides a guide for modifying models to achieve consistency required for integration. The importance of model integration for policy analysis was also reviewed. The integration process was illustrated by reporting the procedures used to incorporate the Hirst model into MEFS.

The work that was begun in integrating the Hirst model into the Mid-Range Energy Demand Forecasting System is continuing both at Oak Ridge (ORNL) and at EIA. The effort at ORNL is designed to improve the internal structure of the model, while the staff at EIA is attempting to improve the interface capability.

In summary, there were five major findings resulting from the task of integrating the Hirst Residential Energy Use Model into the MEFS. First, the use of repeated simulations to generate pseudo data to estimate reduced form energy demand curves is inferior both empirically and computationally to using single pertubations to estimate a narrow range arc elasticity. Second, the MEFS and the Hirst model were developed independently which explains the use of different data concepts in each model. These concepts were reconciled to adjust Hirst to RDFOR. Third, the treatment of abnormal weather conditions and, thus, relation to normal energy consumption patterns must be explicitly incorporated in all updates of the Hirst model. Fourth, the treatment of national gas curtailments in the Hirst model was not only necessary for the simulated comparisons with RDFOR, but served to illustrate the superiority of the Hirst model over RDFOR for use in policy analysis. Finally, the importance of the exogenous driving variables of housing stock, income, and population illustrate the need for improving this sector of the model.

REFERENCES

1. Hirst, Eric and Janet Carney, The ORNL Engineering Economic
    Model of Residential Energy Use, Oak Ridge National Laboratory,
    ORNL/CON 24, July 1978.

2. Elizabeth Chase MacRae, PIES: A User's Guide FEA/N-77/115 June,
    1977 contains a general introduction to the PIES System, which was
    a forerunner of the current Mid-Range Energy Forecasting System.

This paper is a discussion of the validation process
undertaken by the authors. It is not intended nor does
it represent a policy statement of the Department of
Energy or the Electric Power Research Institute.

PANEL SUMMATION

final event of the Workshop was a panel summation that addressed the future
model assessment. The panelists were Martin Greenberger, William Hogan,
ge Lady, David Nissen, Richard Richels, and David Wood. The following is
dited version of the discussion.

DR. WOOD: I would have a few summary comments. First, with all of you,
ve quite enjoyed the last two days and Saul's organization of this Workshop.
things that I already knew have been reaffirmed. For example, whenever
organize an activity such as this, you must be sure to invite David Nissen
Larry Mayer to keep it lively, as well as informative.

issues discussed during the Workshop seemed especially provocative to me.
Weyant's discussion of model assessment versus the Forum, substitutes or
lements, helped me to focus more clearly on what the differences between
e two enterprises might be, and why in the future these apparently separate
vities are likely to merge.

ought Bud Cherry's observation yesterday that much of what goes on in the
m process is not reportable in the traditional documentation and, in fact,
rnalized in the alumnae of that process, is suggestive for further organiza-
al initiative. If the specific product of the Forum is policy research
ies, then the value added is an alumnae sensitized to the uses of policy
ls in policy research.

lly, I want to emphasize a view I have expressed several times during this
shop. Policy model evaluation seems to me to be scientific analysis and
review organized and presented to provide for the information requirements
ll groups--not just modelers--involved in the policy research process.
may appear new or distinctive is the effort to satisfy non-modeler needs
information about models, their scientific validity and applicability to
icular policy issues. As this evaluative aspect of the policy sciences
res, the apparent distinctions between scientific analysis and review and
cy model evaluation will evolve into good scientific practice with whatever
icular methods and practices are appropriate to satisfy the information
s of non-modelers involved in the policy process.

DR. GASS: Thank you Dave. Next, Martin Greenberger.

DR. GREENBERGER: My remarks can be very brief because I spoke a little
earlier today. I would like to join with Dave in congratulating Saul on
iging together an excellent group of people. The National Bureau of Standards
the Department of Energy deserve our thanks for sponsoring this very con-
tive Workshop.

A member of the Workshop told me he felt the future of model assess-
ment hinged on whether modeling is more like writing or physics. What he
meant was that if it is like writing, then just as writing is going to
continue to have its critics for a long time, so modeling will continue t
have its model analysts.

But in physics, he did not see the analogy. I pointed out to him
that there is in face a parallel in physics. There is the theoretical
physicist, who corresponds to the modeler, and then there is the experi-
mental physicist--the fellow in the laboratory testing the theories--who
corresponds to the model analyst. In physics, many more theories are
proposed than survive because of the efforts of the experimentalists.

The question is, can you have the same person serving both functions
In physics, the answer is generally no. Theoretical and experimental
physicists tend to be different people. They work cooperatively, if not
always harmoniously, and there is a very productive symbiosis between
them. It seems to me that it is entirely possible for the same kind of
symbiosis to develop as between model developers and model analysts for
the benefit of the policy-makers and model users.

I am very much encouraged by what has taken place over the past four
years. The field of policy modeling four years ago revealed a very clear
deficiency which it has begun to fill with the development of model
analysis. There is still a way to go, but now we are not asking the
question, "What can we do?" but "What form will it take?" That is encour-
aging.

DR. GASS: Thank you Martin. Next, Bill Hogan.

DR. HOGAN: Well, the first thing I would like to do is to ask
Larry Mayer if he has any plans for dinner? I don't have to catch a
plane until 8 o'clock, so we can see if we can arrange something later on
This is to disprove his statements about his ostracism. So I told him
privately his remarks are very provocative, but although I think almost
everything he said as a factual matter is correct, the general thrust of
his comments was not in keeping with what I viewed as the pleasant
progress of the discussions of this workshop. That progress has been in
laying out the heterogenous nature of models in use and the different
kinds of applications and the different kinds of uses.

I won't repeat all of the taxonomies that were proposed. There were
many and they are developing. If I may use an analogy we are trying to
make the distinction here between the processes that physicists are involv
in vs. the processes that lawyers are involved in. This is the distinctio
between policy research and policy analysis models. We are concerned abou
people using the rules of the physicist in a lawyer's game or using the
practices of lawyers to claim the benefits of physics. We don't want to
confuse these two activities.

I did ask Larry about this and he agreed. I think making the distinction is an important component of what we are doing.

That also leads me to another dimension. As we are developing a more sophisticated view of what we mean by modeling, the modeling process and modeling assessments, we want to make sure that we keep our standards in mind, particularly when we are talking about the end of the spectrum that I refer to as the lawyer's view, i.e., the policy analysis process. I don't know if that analogy is going to hold up, but it certainly is true that the absolute standards that we might appeal to for the scientific evaluation are not going to be relevant for the lawyer.

I was happy to see that a lot of the discussion, at least formally (for example, in the paper by Fred Murphy and Harvey Greenberg), was about validity as a relative concept: this model is more valid than that model; this component is more valid than that component. For the policy analysis purpose, I think that is the only useful piece of information. If you are confronted with a situation where you must make a decision, you want to know how useful the model might be compared with something else that you might have to use in its absence.

The implications of this are many, but I think the most important one in the short run is that we are not very close to being able to specify standards for modeling or model assessments; standards in the rigid sense of being able to give grades and to have necessary conditions, and so forth, for the use of some models. At least this is true in the spectrum of policy analysis. We should continue the kind of work that obviously has been indicated in the discussion here.

In terms of my preferences, I don't know what the future is going to be like, but my preferences for the future would be that there be much more creative energies in trying to understand the end of the spectrum concerned with the use of models, how decisions are influenced by analysis, no matter how formal that analysis may tend to be. That is going to get us into areas of behavioral research. There is some work going on in that. Not all people think about problems the same way, so it is probably true that not all models are going to influence all people the same way and we ought to develop a better understanding of these differences.

The discipline of empirical tests is useful, not only at the scientific end, but also at the policy analysis end. It is an uncomfortable discipline. I endorse the view that we should try to test our models, our concepts, and our ideas at every opportunity. Sometimes those opportunities may be difficult to create, but that is an important element that should be continued.

And then, if we talk about the models and the model use process, we ought to have more information from our own experiences, not from the side of decision makers, but from the side of the modelers about how models are actually used. I would like to see more papers written on the applications of good models and how they are, in fact, applied. I have a

particular interest in this--I recently assumed an area editorship for
Operations Research on energy and environmental problems, and I would be
very anxious to get good papers on the applications of models in discussing
empirical tests, how the models are used, what contributions have been
made, and so forth. We need to publicize that kind of information in
addition to the theoretical descriptions of models that we find so much
easier to write and much easier to criticize.

DR. GASS: Thank you Bill. Next, George Lady.

MR. LADY: I really will be brief as I view my role here more as a
customer than as a contributor and, as a customer, I want to first say I
feel I have been well served, thank Saul Gass for his efforts, and thank
everyone else. I thought it was a very good program. It is very ambitious
to have a program that lasted this long, for two days in a row, and still
have this many people here.

MR. RICHELS: I think they are all getting rides back home with the
people that stayed!

MR. LADY: Anyway, I stayed! So, I will be brief, but let me mention
a few things that I think are important.

First, I liked Martin's chart that organized the who and the why
part of assessments. I am impressed and believe, and others have said it
different ways, that the major first order impact of the process that we
are in will be felt in the upper left hand corner of that particular
organization. That is, what we are really doing is we are in the process
of changing what "modelers normally do" and, in the end, a lot of the
so-called third-party activities and associations, with the ideas we have
been talking about, will disappear. In the end, I think that the modelers
will behave differently than they have so far due to the process that we
are now experiencing.

In my own mind, a lot of the ambiguity that still exists--in terms of
the technical issues and different ways of looking at things--will go away
if scientific principle survives all of this. Model results and things
associated with evaluating model results must be reproducible in general.
That was not talked about too much, but I think that is very critical. I
believe it will be true, far more so in the future than has been in the
past, that the systems that are used to support decisions are going to be
in some sense tractable and available to anyone who wishes to examine the
process that led to the information that formed the decision. This even
extends to the technical issue of model portability.

There have been estimates that assessing a model more or less takes
as many resources as developing the model. I have no basis for questioning
that. Mike Shaw says documentation takes about 25 percent of the resources
to develop a model. This means that, in general, it costs more than twice
as much to do what we have been doing since it is agreed that we have not
really documented or assessed up to the level that we are proposing. It

550

is very expensive. We should emphasize to people that are going to buy
the services that, to do it right, and we have been instructed by the
users to do it right, that it is very expensive.

DR. WOOD: Could I just mention a footnote to that. In production, in
general, the initial production cost is typically just a very small frac-
tion of the total cost of commercializing a product, bringing it to market,
and so forth. I think there is a parallel here, so maybe we should not
be so surprised that something like assessment that makes the model more
usable, more understandable, is going to be expensive and the balance
between the cost of assessing the model and the cost of producing it may
not be excessive at all.

DR. GASS: Thank you George. Next, Dave Nissen.

DR. NISSEN: I would like to thank Saul Gass, the National Bureau of
Standards, and George Lady for making this workshop possible. It has been
a very exciting workshop for me and I am pleasantly surprised at how much
better we understand our ignorance in this area now than we did two days
ago.

Unfortunately I feel compelled to conclude here on a down note. What
I have gathered here is that we do not, either collectively or individually,
understand the role of science in policy modeling very well at all in any
way that is operational. I have two examples of this and, then, a hint of
the reason. One example is that I found myself nodding very interestedly
at David Freedman's description of the READ model. I don't want to get
into the merits of the detailed READ model assessment, but I was saying
to myself, "Gee, it would be terrific to have a model like that." You
could fix some of the things that David said was wrong with it, and I
would conjecture that READ would look a lot like other models. It would
suffer about the same degree of disability and shakiness, it would be
estimated on about the same kinds of data bases, but people would find
such a model cast in a consistent accounting framework to be fabulously
useful. I could make a long list of things for which you might want to
have that kind of regional economic model disaggregation--environmental
assessments, for example, at the air quality control region in analysis,
but aggregated to state-level impacts for reporting. (You need a complete
accounting of economic activity at a fine level of regional disaggregation
to understand the conversion of pollutants to pollution).

But David Freedman concluded that, as a professional statistician,
there was sufficient reason to not build the model essentially because
of econometric problems. It occurred to me that my view of what one did
with models had to be very different from his and that I and my other
colleagues in policy analysis had done a very bad job at communicating
how policy models are actually used, what you do when you build them, and
what they turn out to be valuable for in the policy-making process.

Second, although once again it is charming, I want to share Bill Hogan's view that Larry Meyer's characterization of the role of science in modeling is fun and provocative but somehow it relates to real problems of the use of science in modeling in the same way that Lenny Bruce's dialogue, "Father Flotsky's Triumph," relates to the problems of Vatican II. Larry's comments may be an important counterbalance but they don't speak to the actual problems as they actually arise.

I want to suggest that people who are scientists worry about problems which are much more elegant than people who are policy analysts. That is a commonplace but it is a tremendously important commonplace because it says that what is valuable about a policy model is not the parts that are at the frontier of the professional scientist's concern. (This is not to say there shouldn't be concern over the validity of representation of the hard parts. But usually if the science is uncertain, the responsible policy modeler does scenario variations to box in the uncertainty.) The real contribution of a model like PIES has been in the accounting framework and the consistent integration of behavioral and institutional interactions.

Such a framework for example forces operationality in the policy spec-ification process. We always had a saying that if you couldn't explain a policy to PIES, you probably couldn't write the regulations to enforce it in the real world. I can remember a very long and tedious set of discus-sions we had with the White House people about how to model natural gas regulations—remember this when they still wanted to control gas at a BTU equilibrium price with something. We were trying to model that policy and we said, "What do you want a BTU equilibrium price with?" And they said, "Oil," and we said, "That is not enough; oil where, what kind of oil, crude, resid, distillate? What do you do about quality differentials, location rents, measurement points, and all that kind of stuff?" They finally picked a notional oil control category which literally was designed to give a BTU-equivalent price of $1.75/mcf since they had concluded that $1.75/mcf was the politically saleable price. The use of PIES in this context had almost nothing to do with bias or elasticities or anything else. It had to do with the most rudimentary notion of making operational what you wanted to have happen. (It also happened to illuminate some of the difficulties in the seemingly simple concept of BTU-equivalence as a basis for price regulation.)

I think that until people who are practicing scientists have a richer feel for this, their insights or prescriptions will not be readily used by people using models for policy analysis. On the other hand, that isn't to say the standards of science shouldn't be imposed on the policy analysis process, but, even something as simple as requiring reproducibility must start by making sure the code runs the same on two different machines, which is expensive.

To give an example of how reproducibility questions get hard, consider that all energy system models in which investment occurs embody some implicit or explicit theory about expectations of future prices.

In testing these models against history, you run into serious scientific questions about how to measure indirectly the status of the expectations which are supposed to generate historical behavior. The notions of scientific validity which we inherit from experimental sciences turn out to be very hard to apply in practice to policy modeling, the way models are really used. That is a specific instance of a theme that ran through the history of science part of my talk this morning--that notions of truth and validity are conditional and don't define within themselves their own limits of applicability, that answers about how things ought to be done are by their nature not nearly as satisfactory as we want them to be.

For this reason I am all for pushing through discussions of assessment and validation to something very specific--pick a set of clients and see if what you are going to do satisfies them--like the Congress and the unassisted reader of George Lady's book.

DR. GASS: Thank you Dave. Next, Rich Richels.

DR. RICHELS: I just have a few additional points. I, like my co-panelists, would like to thank you, Saul. I found this a very enjoyable experience and I think I have learned a lot.

There is the idea that has come up several times that the emphasis on model assessment will probably fade. Perhaps modelers will develop more efficient feedback mechanisms and perhaps third party model assessment is not essential for that aspect of the model development process.

I feel, from the user's point of view and as far as the intelligent use of models is concerned, that third party model assessment is essential. I know at EPRI, from the little experience that we have had in this area, we have become convinced that third party model assessment has an awful lot to offer. Although we realize it is a very expensive process, we also realize that it substantially increases the value of the model.

I remember, in the early days when we were starting the PIES, I guess about a year-and-a-half ago, we were looking around for potential modelers. A necessary condition for successful model assessment is the full cooperation of the modeler and we certainly didn't realize how true that was at that time but--well, we did not appreciate it as much back then as we do now.

We are very grateful to Marty Baughman for his cooperation and participation, as far as I am concerned, to an amazing degree. But I think that the payoff is going to be enormous to Marty over the long run. I know that when it is a choice between the Baughman-Joskow model and a model with similar capabilities, we are going to choose the Baughman-Joskow model because we feel comfortable with it right now. It has been assessed, we feel that we understand it, we understand its strength and its weaknesses, and feel that it is a good model. I think that, ultimately, the model assessment process will turn out to be as beneficial to the modelers, as far as promoting his product, as it is to the users. So, I feel very good about that aspect of it.

553

I had a conversation with Dave Wood today about necessary conditions for assessment. He was discussing some thoughts, some problems they are having with the ICF assessment. I said, "I thought we had identified all of the necessary conditions with the Baughman-Joskow assessment and here we have a whole new list of problems that we are having to confront." I think the model assessment process is going to become much more efficient over time as we better understand what are the necessary conditions for assessment. A necessary condition, for example, is a certain level of documentation that will facilitate the process. I think it is going to become incumbent upon the model's sponsors in their statements of work through final model development to make sure that the developmental process and the assessment process can come together. We certainly do not have that right now. I know, from my point of view in being involved in sponsoring future work, that is something I am certainly going to keep in mind.

DR. GASS: Thank you Rich. Before I open it up to the audience, does anybody on the Panel want to make a comment on the comments? Okay. Any questions or comments? Yes, Lincoln?

DR. MOSES: The transactions here for the last couple of days have been very interesting and far ranging. There was an extremal point that appeared today--there have been several extremal points at different times. They have to be different or they would not be extremal I guess. I would like a comment from probably Dave Nissen, but maybe anybody on the Panel. If I understood David Freedman, he at one point said that there may be circumstances where we should say, "We do not know enough to make a model." That is an extremal point I think. I think I heard Dave Nissen say earlier, in an incidental way, that one of the nice things about energy economics is that it is simple enough for us to model it. A superficial reading of the two comments is that Dave Nissen agrees in principle with David Freedman and all there is is some quantitative issues to be resolved, but I am not sure I have it right. Since it is an extremal point, I thought maybe it would be worth a little discussion from the Panel.

DR. GASS: That is an interesting characterization. Dave, do you want to comment?

DR. NISSEN: I think it is a difference in degree. I am more cheered by the quality of the data compared to other data that I have seen used in a way which I regard it as profitable and useful than David Freedman. So that that is a qualitative difference in our assessments of its usefulness.

The idea that a model could get you down to a level where environmental modeling could become operational at an air quality control district, where the physics of an environmental impact start to actually be applicable, seems to me to be a matter of such profound importance to national policy as a whole that even poor models, which could serve as a paradigm for proper data development in this particular area have a very upside potential. It is just that we have to get past the SEAS model

view of the world in the environmental modeling area before we can get anywhere. Everybody runs up to the thing and says, "Well, of course, the dominant issues are environmental," SEAS models at the national level. Knowing how many tons of NOx are dumped in the air is an issue: you do not know what that means about pollution, you do not know what it means about pollutants. So that my responses were based on differences in degree. I think that David Freedman's perception of the usefulness of models in this particular area differs from my perception. The differences in perception reflect the fact that my view of what matters is not, I believe, commonly held. That is a failure that I am contesting.

DR. GASS: Yes? Would you use the microphone?

MR. GRAVES: Joe Graves from Resource Planning Associates. Lest I give the wrong impression, I would like to add my accolades to those of you who have been praising the program. I enjoyed it very much over the last two days. But there is one thing that seems to come to mind to me very frequently in the discussion that was paid very little attention here, and that really is the issue of what goes into the model.

If we step back for a minute and we say the purpose of the effort to validate and to assess is to make it more likely that our consumers will accept what we give them and make us more comfortable about giving them the product, then we also need to be concerned about what is going in and really from two standpoints.

One is the right number, if there is such a thing, and the other is the idea of the standardized number so that, when comparisons are made from different models, that there is some notion of being able to look at different kinds of forecasts, different results from different programs and compare them in a way that makes some kind of sense. I think that it is an issue that needs to have some attention paid to it.

DR. GASS: Thank you. Any comments? Yes, David?

DR. FREEDMAN: I want to make two points for comments perhaps from Dave Nissen.

This morning, David, you told us that one of the great charms, if I understood you correctly, of econometric models was the elegance and rigor of the way in which the coefficients were derived, or fitted, from the data. I guess the point I want to raise is that, if you actually look at the READ model, and perhaps other models, if you actually look at the equations and you actually look at the data, you see that the assumptions behind the fitting procedures, by which I take it you meant something like regression, ordinary least squares, two stage least squares, whatever, you see that the assumptions behind that mathematical theory are violated in virtually every respect. That is one point.

555

And then another point--I want to quote a previous speaker who said that the physical paradigm only goes so far in policy modeling. I guess to an insider, like myself, the thing of it is that once that physical paradigm stops what you start getting into is magic and allocated data, and all kinds of things like that. For an outsider, that seriously diminishes the credibility of the enterprise.

DR. GASS: Any Panel comment on that? Larry?

DR. MAYER: I appreciate the fact that my talk was entertaining. I was a little dismayed that some of the responses were, once again, not by the Panel but by people outside. I am a non-economist throwing bricks at the economists.

I intended to be more than entertaining so what I would like to do was to read a little something which is written by an economist. This is published in "Datalist" which is considered probably the most scholarly journal. It is by a senior economist who is an econometrician, who is concerned with econometric modeling and the current health of econometric modeling and is a full professor at one of our prestigious universities. I will just read a bit and if you want to throw me out at any time, please do.

"Initially econometric models were supposed to test whether a clearly specified theory could be statistically verified, but conclusive tests did not prove to be possible. Models proved not to be up to the task. It has proved to be possible to build many models that are equally statistically from a number of different perspectives. Theories cannot be accepted or rejected based on data; equations cannot stand up over time. The models look solid and precise but they are, in fact, elastic. The data simply are not powerful enough to test and choose among theories.

"Econometrics have shifted from being a tool for testing theories to being a tool for exhibiting theories. It has become a descriptive language rather than a scientific tool. Statistical models are built to show that particular theories are consistent with data and only occasionally can a theory be rejected because of data. As a result, good economic theory is stronger than data, at least in the mind of economists, and, therefore, theory must be imposed on data.

"What started out as being a technique for elevating data relative to theory has ended up doing just exactly the opposite. The theory is never challenged by data and, therefore, never has to be rethought because it is found to be empirically wanting."

Thank you.

DR. GASS: Harvey?

DR. GREENBERG: I just want to --

DR. GREENBERGER: Can I ask a question before you begin?

DR. GASS: The Panel has priority, Harvey.

DR. GREENBERGER: When you look at the dilemma that the economics profession faces today trying to understand and explain a stagnant economy that suffers simultaneously from unemployment and inflation, isn't that an example of data destroying theories?

DR. GASS: Harvey Greenberg.

DR. GREENBERG: I wanted to respond to David Freedman's comment. One of the key assumptions that I assume he is alluding to and what has been violated has to do with correlation of right-hand side variables. The theory says that if you assume this and do this, then this is what you conclude about confidence, and so on. It means that you have a sufficient condition for some of the inferences you want to make statistically. It does not mean that, if they are violated, the thing is bad. Now, I have done a variety of statistical modeling and fore-casting, like in health care and other places, and I do not know of any instance where all of the right-hand side variables are statistically independent and, yet, many of these models are, in fact, useful. And I do not think the judge of usefulness comes from whether you satisfy the sufficient conditions for statistical theory to be valid.

DR. GASS: Thank you, Harvey. Yes, please?

MR. WOOD: Tom Wood from GAO. I have a question, or comment actually. Something has been gnawing on me listening to the question of validation and the purposes of putting forth these models. In a sense, we are trying to provide people better information upon which to make better decisions. The implicit thing being that they did not know as much in the past, so they made bad decisions or, shall we say, not optimum decisions.

I sort of wonder then, if in trying to validate, that modeling isn't approaching some sort of Heisenberg's Uncertainty Principle. If you validate from the past with the "bad decisions," attempting to extrapolate them when you are providing "better information," then it is not directly extrapolatable. So, in other words, the question is, if you are attempting to model a decision process, and by attempting to use as a basis how people made decisions in the past, if those decisions were made badly or not as perfectly as they could be with our great models, then haven't we broken down the ability to use the past?

Again, as Heisenberg said, you cannot specify position and momentum at the same time; at least, if you specify one very well, you lose knowledge of the other. And I sort of question, then, the question of validation. Are we trying to model the differences, decisions, or maybe in the end we are, if we want to backfit in this generic sense, modeling flows rather than decisions?

DR. GASS: Thank you. Does any physicist want to make a comment on that?

MR. JOEL: Saul, can I handle a --

DR. GASS: How could I say no? If you use the microphone, Lambert, and keep it to a couple of minutes.

MR. JOEL: This is going to be less than 30 seconds. Look, the point is, you do not make a decision just once and then go away and God will destroy the Universe in a thunderclap. The idea of giving policymakers slightly better information is that it is not merely that they haven't been able to make optimal decisions in the past. The more nearly good their decisions are, the less frequently they are going to have to change them and, if they have got better information, they are just going to have to do this with less frequency. That is all.

DR. GASS: Thank you, Lambert. I would like to go for just a couple more questions and we will then stop. I know they cut off the heat, but they will put off the lights pretty soon. First, does the Panel have a comment on that? Yes, Dave?

DR. NISSEN: I wanted very briefly to respond to the Heisenberg question. The problem is even more extreme than that. With a proper model of social behavior, we know that government behavior is entirely endogenous and, therefore, we can conclude a fortiori that the entire effect of government policy net is nil.

DR. GASS: Harvey?

DR. WAGNER: I, too, purely as a member of the audience, would like to congratulate you and the others on how fine the conference was in terms of its being comprehensive and thought-provoking.

My interest here, if I may share it with you, is to try to get some perspective on what all of this issue is about because that is fairly new to me, at least in the field of energy, and, if you will permit it, let me share a couple of thoughts of perspective.

The first one is that, for most of us that have been here at the conference, we do have a certain comparative advantage. The comparative advantage, it seems to me from those of you whom I know, is in model building and policy analysis. Another aspect, in terms of impact of models and policy analysis--the politics of it and so on--as important as those aspects are, they really do not play to our own comparative advantage. The net of all of that for me is that we should really give our major emphasis to having better models and better approaches. Other people, who are concerned with some of these other aspects and have a comparative advantage will inevitably pick up those themes. To the extent that we have to sort out our time priorities and our money priorities, we ought to sort them out in a way that we feel best gives a chance to improve models and to improve analysis.

That leads to the second point. I make this because, obviously, the conference was sponsored in part by DOE and we have EPRI here, as well as several other institutions that are concerned with this kind of research. It seems to me that, given a number of comments that were made, both positive, as well as cautious, that, for the time being, that is for the next few years, the institutions involved here should do everything that they are capable of doing to further lots of model-building efforts, and to try to home in on one or two or three and to make them perfect. The best thing that could come out of all of this is the competition among scientists and model builders for approaches and ideas on how to handle these problems. It is really too early to home in.

DR. GASS: Thank you, Harvey.

SPEAKER: He is right!

MR. EVERETT (DOE): I find myself to be mainly a user of models but, also, unfortunately, a caretaker of models that either people on the Panel left me with or certainly other people in this room. At this point, I know the budgetary process is going to be somewhat less than heartening over the next few years and, given that I, myself, and certainly the people within EIA that do most of the analysis of modeling have 50 or 60 of these beasts. How on earth should we choose, given the meager funds that likely will come to this project, which ones we are going to dissect, which should come first? The READ model certainly seems to be a very big target at this particular point but, what next? I would like an answer from the Panel on this one.

DR. LADY: Do you want an answer from me, Charles? Assuming that it doesn't take too many more years to figure out what to do, which may be a very brave assumption, it seems sensible to expect that a reasonable approximation of many of the good ideas that have come up today can be completed on the cycle of model development which will be different. Depending upon what you are talking about, something on the order of three or four years. That is an answer. Is that an answer to your question?

DR. GASS: Charlie, you were concerned about the models in being right now, I gather.

DR. WAGNER: One problem I have is there are models that exist that have been used for forecasting and we put our names next to the forecasts, we publish them, and some of those are going to be replaced. Where there was one, there may be three models in a year. Why don't you pick a model that is an embryo at this point and, before we use it and forecast with it, validate it, rather than something that will be entered into the record as perhaps a bad experiment?

DR. LADY: I think that is the idea, but we have to know what to do. Given that we know what to do, or at least have agreement on some things to do, the idea is to embody it in the model development process.

DR. RICHELS: In the case of our first assessment, it was a model that we could assess. If the documentation is not there or if you do not have the cooperation of the modeler, you might as well forget the assessment, at least at this stage of the game.

Secondly, it is the value of information. What is the model being used for? Is it being used for important policy decisions? That is where we find the greatest need for assessment.

DR. GASS: Any other comments from the Panel? Well, I personally would like -- yes, please?

DR. GLASSEY: I am hearing about the strategies here. Every model developer that is currently under contract to the EIA to do models, must have his models assessed before we pay him.

DR. GASS: That is if we can set the ground rules. Alan, would you like to make a comment? Alan Goldman.

DR. GOLDMAN: Two very quick remarks. One of them is a distinctly self-serving suggestion to the Chairman. Some of us may have some reactions to this meeting which we were unable so quickly to articulate, or will not articulate now because of the lateness of the hour. Perhaps you might care to declare the proceedings open to late submissions to these remarks.

DR. GASS: Yes, that is definitely true. The deadline is March 31.

DR. GOLDMAN: Okay. My second comment is again as representative of the host institution to thank you for the quality of your discussions from the floor, delivered papers, and zitzfleisch.

DR. GASS: Thank you very much. I would like to thank the Panel, both as a Panel and speakers. I would like to thank the other speakers, and I really would like to thank this tenacious audience for staying with us. Thank you very, very much.

WORKSHOP ON

VALIDATION AND ASSESSMENT ISSUES OF
ENERGY MODELS

January 10 - 11, 1979
Lecture Room A, Main Administration Building
National Bureau of Standards
Gaithersburg, Maryland

PROGRAM

January 10, 1979 (Wednesday)

9:00 - 9:10...Welcome...........................Saul I. Gass
Chairman
U. of MD/NBS

Alan J. Goldman
NBS

Roger Glassey
DOE

9:10 - 9:20...Introductory Remarks...............Lincoln Moses, DOE

9:20 - 9:45...Model Assessment and Validation:
Issues, Structure, and EIA Program
Goals..............................George Lady, DOE

9:45 - 10:15...Model Assessment and the Policy
Research Process: Current Practice
and Future Promise.................David O. Wood, MIT

10:15 - 10:30...Discussant.......................William W. Hogan,
Harvard

10:30 - 10:45...Coffee

10:45 - 11:15...The Energy Modeling Forum..........James Sweeney,
Stanford

11:15 - 11:45...Electric Load Forecasting: Probing
the Issues with Models.............Bernard H. Cherry,
General Public
Utilities

11:45 - 12:15...Assessing the ICF Coal and Electric
Utilities Model....................Neil L. Goldman, MIT
James Gruhl, MIT

12:15 - 12:45...Validation and Assessment of ICF's
Coal and Electric Utilities Model..C. Hoff Stauffer, Jr.,
ICF

12:45 - 1:30...Lunch

1:30 - 2:00...What Do We Mean When We Say
Validation.........................Peter W. House, DOE
Richard Ball, DOE

2:00 - 3:00...Third Party Model Assessment.......Richard Richels, EPRI
David Kresge, MIT

3:00 - 3:15...Coffee

3:15 - 3:45...Reflections on the Model Assessment
Process: A Modeler's Perspective..Martin L. Baughman,
U. of Texas

3:45 - 4:15...The Texas National Energy Modeling
Project and Evaluation of EIA's
Energy Midrange Forecasting Model..Milton Holloway, TEAC

4:15 - 4:45...Assessment of the Midterm Electric
Utility Submodel...................Fred Murphy, DOE

4:45 - 5:00...Model Management Issues............Saul I. Gass,
U. of MD/NBS

January 11, 1979 (Thursday)

9:00 - 9:30...Impacts of Assessment on The Modeling
Process............................David Nissen,
Chase Manhattan

9:30 - 10:00...The Energy Modeling Forum and Model
Assessments: Substitutes or Comple-
ments.............................John Weyant,
Stanford

10:00 - 10:30...A Way of Thinking About Model
Analysis..........................Martin Greenberger,
Johns Hopkins

10:30 - 10:45...Coffee

10:45 - 11:15...Energy Modeling Methods and Related
Validation Issues...................Edward Cazalet,
DFI

11:15 - 12:15...Energy Modeling Methods and Related
Validation Issues (Continued).........Shail C. Parikh,
Stanford

William Marcuse,
Brookhaven

Thomas Sparrow,
Purdue

12:15 - 12:35...Model Access and Documentation........Michael Shaw, LMI

12:35 - 1:00...Assessing the Regional Energy Activity
and Demographic Model................David Freedman,
Berkeley

1:00 - 1:45...Lunch

1:45 - 2:15...Assessment and Selection of Models
for Econometric Analysis..............Edward A. Hudson,
DRI

Dale W. Jorgenson,
Harvard

2:15 - 2:45...Econometric Model Forecasting and
Related Validation Issues.............Edwin Kuh, MIT

2:45 - 3:00...Coffee

3:00 - 3:30...Assessing Energy Methodologies and
Models: Evaluating Policy Process
Analyses...........................Lawrence Mayer,
Princeton

3:30 - 3:55...Sensitivity Analysis and Validation
of Coupled Models....................Fred Schweppe, MIT

3:55 - 4:20...Interactive System for Diagnostic
Analysis...........................Harvey Greenberg,
DOE

4:20 - 4:45...Integration of the Structural Resi-
dential Energy Use Model into the
Mid-Range Energy Forecasting System...Frank Hopkins, DOE

Lewis Rubin, DOE

4:45 - 5:30...The Future of Assessment (Panel
Summation)..........................Martin Greenberger,
Johns Hopkins

William Hogan,
Harvard

George Lady, DOE

David Nissen,
Chase Manhattan

Richard Richels,
EPRI

562

David O. Wood, MIT

WORKSHOP ON VALIDATION AND ASSESSMENT OF ENERGY MODELS

January 10 - 11, 1979

Dr. Norman Agin
Mathtech, Inc.
P. O. Box 2392
Princeton, NJ 08540

Dr. John Andelin
Senate Committee on Science and
  Technology
U. S. Congress
Washington, DC 20510

Mr. Robert Andros
104 N. Harrison Road
Sterling, VA 22170

Prof. Benjamin Avi-Itzhak
Dept. of Operations Research
Stanford University
Stanford, CA 93705

Dr. Egon Balas
Carnegie Mellon University
Pittsburgh, PA 15213

Dr. Richard Ball
Office of Technology Impacts
Department of Energy
Washington, DC 20545

Ms. Mary Barcella
Logistics Management Institute
4701 Sangamore Road
Washington, DC 20016

Mr. James P. Barnett
National Bureau of Standards
Center for Building Technology
Bldg. 226, Rm. B104
Washington, DC 20234

Prof. Richard Bassman
Department of Economics
Texas A&M
College Station, TX 77843

Dr. Martin L. Baughman
Center for Energy Studies
University of Texas at Austin
Austin, TX 78712

Dr. Robert L. Bivins
L. A. S. L.
Mail Stop 606
Los Alamos, NM 87545

Dr. David Brillinger
Department of Statistics
University of California
Berkeley, CA 94720

Dr. Horace Brock
SRI
Menlo Park, CA 94025

Mr. Frank Capece
12015 William & Mary Circle
Woodbridge, VA 22192

Dr. Edward G. Cazalet
Decision Focus, Inc.
1801 Page Mill Road
Palo Alto, CA 94304

Ms. Ellen Cherniavsky
Brookhaven National Laboratory
Upton, NY 11973

Mr. Bernard Cherry
General Public Utilities
260 Cherry Hill Road
Parsippany, NJ 07054

Mr. Phil Childress
Department of Energy
Room 8433
12th & Pennsylvania Ave.
Washington, DC 20461

Mr. Wallace Cohen
U. S. GAO
441 G Street, NW
Washington, DC 20548

Mr. Charles B. Colton
Systems Consultants, Inc.
1054 31st Street, NW
Washington, DC 20007

Dr. Burton H. Colvin
National Bureau of Standards
Center for Applied Mathematics
Washington, DC 20234

Dr. Leon Cooper
Southern Methodist University
Dallas, TX 75222

Dr. Robert Crosby
DPB25, DOT
400 7th Stteet, SW
Washington, DC 20590

Mr. James T. Crowell
c/o Systems Consultants, Inc.
1054 31st Street, NW
Washington, DC 20007

Dr. George B. Dantzig
Dept. of Operations Research
Stanford University
Stanford, CA 94305

Mr. Kenneth G. Darrow, Jr.
Gas Research Institute
10 W. 35th Street
Chicago, IL 60616

Mr. Alan D. Davies
National Bureau of Standards
Center for Consumer Product Technology
Bldg. 220, Rm. A359
Washington, DC 20234

Dr. Robert Dial
UMTA UPM 20
Room 9311
Department of Transportation
Washington, DC 20590

Dr. R. Drenick
Dept. of Electrical Engineering
Polytechnic Institute of New York
333 Jay Street
Brooklyn, NY 11201

Mr. Charles Everett
U. S. Department of Energy
Room 4447, Federal Bldg.
12th & Pennsylvania Ave., NW
Room 4530
Washington, DC 20461

Mr. Robert Eynon
Department of Energy
1200 Pennsylvania Ave., NW
Room 4530
Washington, DC 20461

Mr. Richard Farman
EG&G Idaho, Inc.
Box 1625
Idaho Falls, ID 83401

Ms. Anna Fletcher
Booz-Allen & Hamilton
4330 East West Highway
Bethesda, MD 20014

Mr. Andy Ford
Energy Systems and Economic
 Analysis Group
Los Alamos Scientific Lab
Los Alamos, New Mexico 87545

Dr. David Freedman
Statistics Department
University of California
Berkeley, CA 94720

Mr. Ramesh Ganeriwal
Energy & Environmental Analysis Inc.
1111 North 19th Street
Arlington, VA 22209

Dr. Saul I. Gass
National Bureau of Standards
Center for Applied Mathematics
Washington, DC 20234

Ms. Phyllis Gilmore
Department of Energy
12th and Pennsylvania Ave., NW
Room 4530
Washington, DC 20461

Dr. C. Roger Glassey
Office of Applied Analysis
Energy Information Administration
12th & Pennsylvania Ave., NW
Washington, DC 20461

Dr. Richard Goettle
Chase Manhattan Bank
1 Chase Manhattan Plaza
New York, NY 10015

Dr. Alan J. Goldman
Center for Applied Mathematics
National Bureau of Standards
Washington, DC 20234

Dr. Neil Goldman
Energy Laboratory
E-38, Room 413
MIT
Cambridge, MA 02139

Mr. Joseph Graves
Resource Planning Assoc., Inc.
1901 L Street, NW
Washington, DC 20036

Dr. Harvey Greenberg
Federal Bldg.
1200 Pennsylvania Ave., NW
Room 4522
Washington, DC 20461

Dr. Martin Greenberger
The Johns Hopkins University
Dept. of Mathematical Sciences
Charles & 34th Street
Baltimore, MD 21218

Mr. Frank J. Gross
9702 Shipwright Drive
Burke, VA 22015

Mr. James Gruhl
142 Plain Road
Wayland, MA 01778

Mr. H. N. Hantzes
Code 2032B
Naval Facility Engineering Command
200 Stoval Street
Alexandria, VA 22332

Mr. Larry Hare
National Science Foundation
1800 G Street, NW
Washington, DC 20550

Dr. Carl Harris
Mathtech, Inc.
1611 N. Kent Street, Suite 200
Arlington, VA 22209

Mr. Roy Harvey
Control Analysis Corp.
800 Welch Road
Palo Alto, CA 94304

Mr. J. Michael Hihn
4324 Rowalt Drive, #202
College Park, MD 20740

Mr. David S. Hirshfeld
Ketron, Inc.
1400 Wilson Blvd.
Arlington, VA 22209

Dr. Karla Hoffman
Center for Applied Mathematics
National Bureau of Standards
Washington, DC 20234

Dr. Kenneth C. Hoffman
National Center for Analysis
 of Energy Systems
Brookhaven National Laboratory
Upton, Long Island, NY 11973

Dr. William Hogan
John F. Kennedy School of
 Government
79 Boylston Street
Cambridge, MA 02139

Mr. Milton Holloway
Texas Energy Advisory Council
7703 North Lamar Blvd.
Austin, TX 78752

Dr. Frank Hopkins
Energy Information Admin.
Department of Energy
Washington, DC 20545

Dr. Peter W. House
Office of Technology Impacts
Department of Energy
Washington, DC 20545

Dr. Edward A. Hudson
c/o Dr. D. W. Jorgensen
Department of Economics
Harvard University
Cambridge, MA 02138

Dr. Howard Hung
National Bureau of Standards
Center for Applied Mathematics
Washington, DC 20234

Mr. Richard H. F. Jackson
Center for Applied Mathematics
National Bureau of Standards
Washington, DC 20234

Mr. Allen Jacobs
7A-029 Forrestal Bldg.
1000 Independence Ave., SW
Washington, DC 20585

Dr. Dale W. Jorgenson
Department of Economics
Harvard University
Cambridge, MA 02138

Mr. James Just
DHR, Inc.
Suite 414
1055 Thomas Jefferson Street, NW
Washington, DC 20007

Mr. Richard W. Kelley
Professional Audit Review Team
U.S. GAO
441 G Street, NW
Washington, DC 20548

Dr. J. David Khazzoom
2170 20th Ave.
San Francisco, CA 94116

Mr. Chris Kline
Survey Research Center
Institute for Social Research
Ann Arbor, MI 48108

Mr. David Kline
729 Massachusetts Ave., NE
Washington, DC 20002

Dr. David Knapp
Energy Economics
Chase Manhattan Bank
1 Chase Manhattan Plaza
New York, NY 10015

Dr. David Kresge
53 Church Street
Cambridge, MA 02138

Mr. Bill Kruvant
GAO, Room 5108
451 G Street, NW
Washington, DC 20548

Dr. Edwin Kuh
Center for Computational Research in
 Economics and Management Science
E-38, Room 210, MIT
Cambridge, MA 02139

Mr. William G. Kurator
Department of Energy
Room 4515
12th and Pennsylvania Ave.
Washington, DC 20026

Mr. Andy S. Kydes
NCAES, Bldg. 475
Brookhaven National Laboratory
Upton, NY 11973

Dr. George M. Lady
11307 Soward Drive
Kensington, MD 20795

Dr. Averill M. Law
Department of Industrial Eng.
1513 University Ave.
University of Wisconsin
Madison, WI 53706

Dr. Michael Lerner
EIA
Department of Energy
Washington, DC 20545

Mr. Joel Levy
Center for Building Technology
National Bureau of Standards
Washington, DC 20234

Dr. Dilip Limaye
Synergic Resources Corp.
355 East Gowen Avenue
Philadelphia, PA 19119

Dr. Hubert Lipinski
Institute for the Future
2740 Sand Hill Road
Menlo Park, CA 94025

Dr. Stanley T. Liu
Center for Building Technology
National Bureau of Standards
Washington, DC 20234

Dr. G. S. Maddala
Department of Economics
University of Florida
Gainesville, FL 32611

Dr. Alan Manne
Department of Operations Research
Stanford University
Stanford, CA 94305

Dr. William Marcuse
Bldg. 475
Brookhaven National Laboratory
Upton, NY 11777

Dr. Kneale T. Marshall
Deputy Chief of Naval Operations (MPT)
Code Op-OIT
Department of the Navy
Washington, DC 20370

Dr. Lawrence S. Mayer
210 Fine Hall, Dept. of Statistics
Princeton University
Princeton, NJ 08540

Mr. Michael D. McKay
Los Alamos Scientific Laboratory
S-1, MS 606
Los Alamos, New Mexico 87545

Dr. Dennis L. Meadows
Systems Dynamics Group, Box 8000
Dartmouth College
Hanover, New Hampshire 03755

Ms. Jacquelyn S. Mitchell
5404 Helm Court
Fairfax, VA 22032

Mr. Terry Morlan
EIA
24309 Ridge Road
Damascus, MD 20750

Dr. Lincoln Moses
Administrator, EIA
Department of Energy
Washington, DC 20545

Dr. John Mulvey
School of Engineering & Applied Science
Princeton University
Princeton, NJ 08540

Dr. Frederic H. Murphy
Energy Information Administration
Department of Energy
Washington, DC 20545

Mr. Dale Nesbitt
Decision Focus, Inc.
1801 Page Mill Road
Palo Alto, CA 94304

Dr. David Nissen
Chase Manhattan Bank
1 Chase Manhattan Plaza
New York, NY 10015

Dr. William Nordhaus
Council for Economic Advisors
Executive Office Building
Washington, DC 20005

Mr. Richard P. O'Neill
Applied Analysis/EIA/DOE
12th and Pennsylvania Ave., NW
Washington, DC 20461

Dr. Shailendra C. Parikh
Dept. of Operations Research
Stanford University
Stanford, CA 94305

Dr. Robert Pendley
Environment, Energy Resources
 Group
PRA, Room 1240
National Science Foundation
1800 G Street, NW
Washington, DC 20550

Dr. William Pierskalla
National Health Care Management
 Center
University of Pennsylvania
Philadelphia, PA 19104

Dr. David Pillati
Brookhaven National Laboratory
Upton, Long Island, New York

Mr. Frank Potter
House Subcommittee on Energy
 & Power
U. S. Congress
Washington, DC 20515

Ms. Melinda G. Rackoff
Gas Research Institute
10 W. 35th Street
Chicago, IL 60616

Dr. Paul Randolph
Energy Economics, Chase Manhattan Bank
1 Chase Manhattan Plaza
New York, NY 10015

Dr. Leo Rapoport
Dept. of Geological Sciences
VPI & SU
Blacksburg, VA 24061

Dr. Richard Richels
EPRI
P. O. Box 10412
Palo Alto, CA 94303

Mr. David C. Roberts
University of Michigan
Highway Safety Research Institute
Huron Parkway and Baxter Road
Ann Arbor, MI 48109

Mr. .R. C. Robinson
U. S. Nuclear Regulatory Commission
Mail Stop 1130 S. S.
Washington, DC 20555

Ms. Arlene Rosenbaum
Dept. of Energy
Office of Conservation Planning & Policy
Mail Stop 2221C/ Room 2220
Washington, DC 20545

Mr. Paul Roth
Institute for Computer Sciences and Tech
National Bureau of Standards
Washington, DC 20234

Dr. Lewis Rubin
Energy Information Administration
Department of Energy
Washington, DC 20545

Dr. Rama Sastry
Acting Chief, Applied Analysis Branch
DEI, Mail Stop E-201
Department of Energy
Washington, DC 20545

Ms. Patsy Saunders
Center for Applied Mathematics
National Bureau of Standards
Washington, DC 20234

Mr. Fred C. Schweppe
Room 10-176
MIT
Cambridge, MA 02139

Dr. Michael L. Shaw
Logistics Management Institute
4701 Sangamore Road
Washington, DC 20016

Mr. Thomas P. Sheahen
National Bureau of Standards
Thermal Processes Division
Washington, DC 20234

Dr. John Shewmaker
Office of Information Validation
EIA/DOE
Washington, DC 20545

Dr. Charles Smith
Energy Information Administration
Washington, DC 20545

Dr. A. L. Soyster
270 Whittemore Bldg.
VPI&SU
Blacksburg, VA 24061

Dr. Thomas Sparrow
Department of Industrial Eng.
Purdue University
W. Lafayette, Indiana 47907

Dr. C. Hoff Stauffer, Jr.
ICF, Inc.
Suite 400
1990 M Street, NW
Washington, DC 20036

Dr. Leroy Stewart
AGA
1515 Wilson Blvd.
Arlington, VA 22209

Dr. William Stitt
ICF, Inc.
Suite 400
1990 M Street, NW
Washington, DC 20036

Dr. David Strom
Resource Planning Association
1901 L Street, NW
Washington, DC 20036

Dr. James Sweeney
Energy Modeling Forum
Terman Engineering Center
Stanford University
Stanford, CA 94305

Dr. Katsuaki Terasawa
Jet Propulsion Laboratory
Mail Stop 506-316
4800 Oak Grove Drive
Pasadena, CA 91103

Dr. Bruce W. Thompson
3006 Twisting Lane
Bowie, MD 20715

Dr. Nicolai Timenes, Jr.
Office of the Assistant Secretary for
  Policy & Evaluation
U. S. Department of Energy
Washington, DC 20545

Mr. Leon Tucker
AGA
1515 Wilson Blvd.
Arlington, VA 22209

Ms. Janice H. Varner
B-K Dynamics
15825 Shady Grove Road
Rockville, MD 20850

Prof. Harvey Wagner
Business School
University of North Carolina
Carroll Hall
Chapel Hill, NC 27514

Dr. Richard Waller
Office of PRA
National Science Foundation
1800 G Street, NW
Washington, DC 20550

Dr. John P. Weyant
Energy Modeling Forum
Room 410 Terman Engineering
 Center
Stanford University
Stanford, CA 94305

Dr. William B. Widhelm
9004 Manordale Lane
Ellicott City, MD 21043

Dr. David O. Wood
Energy Laboratory
E-38, Room 417
MIT
Cambridge, MA 02139

Dr. Thomas J. Woods
17 Mills Road
Gaithersburg, MD 20760

Dr. David Yancey
352 West Oak Street
West Lafayette, Indiana 47906

Dr. Oliver Yu
EPRI
P. O. Box 10412
Palo Alto, CA 94303

U.S. DEPT. OF COMM. BIBLIOGRAPHIC DATA SHEET	1. PUBLICATION OR REPORT NO. NBS SP 569	2. Gov't Accession No.	3. Recipient's Accession No.

TITLE AND SUBTITLE	5. Publication Date
Proceedings of the Workshop on Validation and Assessment Issues of Energy Models held January 10-11, 1979 at NBS, Gaithersburg, Maryland	February 1980
	6. Performing Organization Code

AUTHOR(S) Saul I. Gass, Editor	8. Performing Organ. Report No.

PERFORMING ORGANIZATION NAME AND ADDRESS	10. Project/Task/Work Unit No.
NATIONAL BUREAU OF STANDARDS DEPARTMENT OF COMMERCE WASHINGTON, DC 20234	11. Contract/Grant No.

SPONSORING ORGANIZATION NAME AND COMPLETE ADDRESS *(Street, City, State, ZIP)*	13. Type of Report & Period Covered
Operations Research Division Center for Applied Mathematics National Bureau of Standards Washington, DC 20234	14. Sponsoring Agency Code

SUPPLEMENTARY NOTES

Library of Congress Catalog Card Number: 79-600216

☐ Document describes a computer program; SF-185, FIPS Software Summary, is attached.

ABSTRACT *(A 200-word or less factual summary of most significant information. If document includes a significant bibliography or literature survey, mention it here.)*

The Workshop on Validation and Assessment Issues of Energy Models, held at the National Bureau of Standards, Gaithersburg, Maryland (January 10-11, 1979), was funded by the Energy Information Administration of the Department of Energy (DOE), Washington, DC. Organized by the Bureau's Operations Research Division, the Workshop was designed to be a forum in which the theoretical and applied state-of-the-art of validation and assessment, with emphasis on energy models, could be presented and discussed. Speakers addressed the following areas: DOE's activities in assessment and validation, taxonomy and structure of assessment and validation, the relationship between model assessment and policy research, the Electrical Power Research Institute's Energy Modeling Forum and projects, independent third-party model assessment, the Texas National Energy Modeling Project, management and improvement of the modeling process, complexity of model evaluation, definitions and structure of model assessment approaches, model access and documentation, assessment of specific models by the MIT Energy Laboratory and other groups, energy and econometric models, and sensitivity analysis. This volume documents the Proceedings (papers and discussion) of the Workshop.

KEY WORDS *(six to twelve entries; alphabetical order; capitalize only the first letter of the first key word unless a proper name; separated by semicolons)*

Assessment; documentation; econometric models; energy modeling forum; energy models; evaluation; mathematical models; model management; model access; sensitivity analysis; validation.

AVAILABILITY	☒ Unlimited	19. SECURITY CLASS (THIS REPORT)	21. NO. OF PRINTED PAGES
☐ For Official Distribution. Do Not Release to NTIS		UNCLASSIFIED	559
☒ Order From Sup. of Doc., U.S. Government Printing Office, Washington, DC 20402, SD Stock No. 003-003-02155-5		20. SECURITY CLASS (THIS PAGE)	22. Price
☐ Order From National Technical Information Service (NTIS), Springfield, VA. 22161		UNCLASSIFIED	$9.50

NBS TECHNICAL PUBLICATIONS

PERIODICALS

JOURNAL OF RESEARCH—The Journal of Research of the National Bureau of Standards reports NBS research and development in those disciplines of the physical and engineering sciences in which the Bureau is active. These include physics, chemistry, engineering, mathematics, and computer sciences. Papers cover a broad range of subjects, with major emphasis on measurement methodology and the basic technology underlying standardization. Also included from time to time are survey articles on topics closely related to the Bureau's technical and scientific programs. As a special service to subscribers each issue contains complete citations to all recent Bureau publications in both NBS and non-NBS media. Issued six times a year. Annual subscription: domestic $17; foreign $21.25. Single copy, $3 domestic; $3.75 foreign.

NOTE: The Journal was formerly published in two sections: Section A "Physics and Chemistry" and Section B "Mathematical Sciences."

DIMENSIONS/NBS—This monthly magazine is published to inform scientists, engineers, business and industry leaders, teachers, students, and consumers of the latest advances in science and technology, with primary emphasis on work at NBS. The magazine highlights and reviews such issues as energy research, fire protection, building technology, metric conversion, pollution abatement, health and safety, and consumer product performance. In addition, it reports the results of Bureau programs in measurement standards and techniques, properties of matter and materials, engineering standards and services, instrumentation, and automatic data processing. Annual subscription: domestic $11; foreign $13.75.

NONPERIODICALS

Monographs—Major contributions to the technical literature on various subjects related to the Bureau's scientific and technical activities.

Handbooks—Recommended codes of engineering and industrial practice (including safety codes) developed in cooperation with interested industries, professional organizations, and regulatory bodies.

Special Publications—Include proceedings of conferences sponsored by NBS, NBS annual reports, and other special publications appropriate to this grouping such as wall charts, pocket cards, and bibliographies.

Applied Mathematics Series—Mathematical tables, manuals, and studies of special interest to physicists, engineers, chemists, biologists, mathematicians, computer programmers, and others engaged in scientific and technical work.

National Standard Reference Data Series—Provides quantitative data on the physical and chemical properties of materials, compiled from the world's literature and critically evaluated. Developed under a worldwide program coordinated by NBS under the authority of the National Standard Data Act (Public Law 90-396).

NOTE: The principal publication outlet for the foregoing data is the Journal of Physical and Chemical Reference Data (JPCRD) published quarterly for NBS by the American Chemical Society (ACS) and the American Institute of Physics (AIP). Subscriptions, reprints, and supplements available from ACS, 1155 Sixteenth St., NW, Washington, DC 20056.

Building Science Series—Disseminates technical information developed at the Bureau on building materials, components, systems, and whole structures. The series presents research results, test methods, and performance criteria related to the structural and environmental functions and the durability and safety characteristics of building elements and systems.

Technical Notes—Studies or reports which are complete in themselves but restrictive in their treatment of a subject. Analogous to monographs but not so comprehensive in scope or definitive in treatment of the subject area. Often serve as a vehicle for final reports of work performed at NBS under the sponsorship of other government agencies.

Voluntary Product Standards—Developed under procedures published by the Department of Commerce in Part 10, Title 15, of the Code of Federal Regulations. The standards establish nationally recognized requirements for products, and provide all concerned interests with a basis for common understanding of the characteristics of the products. NBS administers this program as a supplement to the activities of the private sector standardizing organizations.

Consumer Information Series—Practical information, based on NBS research and experience, covering areas of interest to the consumer. Easily understandable language and illustrations provide useful background knowledge for shopping in today's technological marketplace.

Order the above NBS publications from: Superintendent of Documents, Government Printing Office, Washington, DC 20402.

Order the following NBS publications—FIPS and NBSIR's—from the National Technical Information Services, Springfield, VA 22161.

Federal Information Processing Standards Publications (FIPS PUB)—Publications in this series collectively constitute the Federal Information Processing Standards Register. The Register serves as the official source of information in the Federal Government regarding standards issued by NBS pursuant to the Federal Property and Administrative Services Act of 1949 as amended, Public Law 89-306 (79 Stat. 1127), and as implemented by Executive Order 11717 (38 FR 12315, dated May 11, 1973) and Part 6 of Title 15 CFR (Code of Federal Regulations).

NBS Interagency Reports (NBSIR)—A special series of interim or final reports on work performed by NBS for outside sponsors (both government and non-government). In general, initial distribution is handled by the sponsor; public distribution is by the National Technical Information Services, Springfield, VA 22161, in paper copy or microfiche form.

BIBLIOGRAPHIC SUBSCRIPTION SERVICES

The following current-awareness and literature-survey bibliographies are issued periodically by the Bureau:

Cryogenic Data Center Current Awareness Service. A literature survey issued biweekly. Annual subscription: domestic $25; foreign $30.

Liquefied Natural Gas. A literature survey issued quarterly. Annual subscription: $20.

Superconducting Devices and Materials. A literature survey issued quarterly. Annual subscription: $30. Please send subscription orders and remittances for the preceding bibliographic services to the National Bureau of Standards, Cryogenic Data Center (736) Boulder, CO 80303.

SPECIAL FOURTH-CLASS RATE
BOOK

1298