Check for updates

NIST Special Publication
NIST SP 1335

A Method for Assessing and Communicating the Quality of Evidence Used in Establishing Community Resilience Frameworks

Michael D. Gerst
Maria K. Dillard

This publication is available free of charge from:
https://doi.org/10.6028/NIST.SP.1335

NIST NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY
U.S. DEPARTMENT OF COMMERCE

NIST Special Publication
NIST SP 1335

A Method for Assessing and Communicating the Quality of Evidence Used in Establishing Community Resilience Frameworks

Michael D. Gerst
Maria K. Dillard
Material and Structural Systems Division
Engineering Laboratory

This publication is available free of charge from:
https://doi.org/10.6028/NIST.SP.1335

June 2025



U.S. Department of Commerce
Howard Lutnick, Secretary

National Institute of Standards and Technology
Craig Burkhardt, Acting Under Secretary of Commerce for Standards and Technology and Acting NIST Director

NIST SP 1335
June 2025

Certain equipment, instruments, software, or materials, commercial or non-commercial, are identified in this paper in order to specify the experimental procedure adequately. Such identification does not imply recommendation or endorsement of any product or service by NIST, nor does it imply that the materials or equipment identified are necessarily the best available for the purpose.

NIST Technical Series Policies
Copyright, Use, and Licensing Statements
NIST Technical Series Publication Identifier Syntax

Publication History
Approved by the NIST Editorial Review Board on 2025-05-12

How to Cite this NIST Technical Series Publication
Gerst MD and Dillard MK (2025) A Method for Assessing and Communicating the Quality of Evidence Used in Establishing Community Resilience Frameworks. (National Institute of Standards and Technology, Gaithersburg, MD), NIST Special Publication (SP) NIST SP 1335. https://doi.org/10.6028/NIST.SP.1335

Author ORCID iDs
Michael D. Gerst: 0000-0002-5281-3228
Maria K. Dillard: 0000-0003-1772-8914

Abstract

Community resilience refers to the concept of a community being able to prepare, respond, recover, and adapt to disruptions in a way that does not compromise long-term community functionality and well-being. While it is not directly measurable, the concept is increasingly used in planning and management. Consequently, creating indicators that can provide information about unmeasurable concepts is a common methodological choice for quantifying resilience and tracking changes due to mitigation actions, investments, and recovery.

Developing community resilience indicators can be challenging because resilience covers many different subject areas and sectors (e.g., infrastructure, health, economics), leading to researchers needing a broad knowledge base to justify indicator choices. One option is to rely on expert panels to provide input, but panels can be resource-intensive and infeasible for many indicator projects. As an alternative, resilience indicator frameworks usually use non-systematic literature reviews to provide theoretical and empirical backing. Relying on this approach can lead to problems in assessing and communicating indicator credibility because non-systematic methods generally lack established procedures for appraising the quality of evidence provided by selected literature; these procedures are especially important in cases where indicators undergo no further construct validation.

This study proposes a method based on Bayesian confirmation theory for appraising and communicating the informativeness of literature used to make indicator inclusion choices. In a test application to commonly suggested resilience indicators, the work illustrates the utility of the method over other options, as it allows for transparent judgment and aggregation of limited data.

Keywords

Indicators; metrics; community resilience; synthesis; expert judgment; quality of evidence.

Table of Contents

1. Introduction **1**
2. Background **3**
3. Framework **6**
3.1. Theoretical Underpinning 6
3.2. Study Types 12
3.2.1. Quantitative Empirical Evidence 12
3.2.2. Qualitative Empirical Evidence 14
3.2.3. Mixed Methods Evidence 15
3.2.4. Synthesis Evidence 16
4. Results **21**
4.1. Test Case 21
4.2. Summary Statistics 21
4.3. Assessing Validated Indicator Studies 24
4.4. Assessing Empirical Studies 28
5. Discussion **34**
6. Conclusion **35**
References **36**

List of Tables

Table 1. Probability table for evidence *E* conditional on a causal hypothesis *S* **7**
Table 2. Probability table for a causal hypothesis *S* conditional on an indicator hypothesis *H* **7**
Table 3. Bayes factor ranges for qualitative evidence strength descriptors. **9**
Table 4. Types of quantitative empirical studies. **12**
Table 5. Types of qualitative empirical studies. **15**
Table 6. Types of synthesis studies. **18**
Table 7. Resilience indicator frameworks included in this study **21**
Table 8. Commonly used indicators and their inclusion in selected frameworks. **22**
Table 9. Application of framework to cited synthesis studies. **26**
Table 10. Application of framework to empirical studies. **30**

List of Figures

Fig. 1. Different types of validity. **3**

Fig. 2. Probability ranges of qualitative probability descriptors. 8

Fig. 3. Bayes factor for combinations of 'error' and 'power' when it is *virtually certain* (*virtually impossible*) an indicator is (not) causally related to resilience, given that a causal relationship exists. Solid lines mark boundaries between adjacent Bayes factor qualitative descriptors. 9

Fig. 4. Bayes factor for combinations of 'error' and 'power' when it is *exceptionally likely* (*exceptionally unlikely*) an indicator is (not) causally related to resilience, given that a causal relationship exists. Solid lines mark boundaries between adjacent Bayes factor qualitative descriptors. 10

Fig. 5. Bayes factor for combinations of 'error' and 'power' when it is *extremely likely* (*extremely unlikely*) an indicator is (not) causally related to resilience, given that a causal relationship exists. Solid lines mark boundaries between adjacent Bayes factor qualitative descriptors. 11

Fig. 6. Bayes factor for combinations of 'error' and 'power' when it is *very likely* (*very unlikely*) an indicator is (not) causally related to resilience, given that a causal relationship exists. Solid lines mark boundaries between adjacent Bayes factor qualitative descriptors. 11

Fig. 7. Mean number of sources cited per indicator and number of indicators for each framework. 23

Fig. 8. Hierarchical representation of included study types. Moving from the center, division of each subsequent level is proportional to the adjacent type in the preceding level. 24

Acknowledgments

The authors are grateful to the reviewers who contributed to the development of this manuscript.

Author Contributions

Gerst: Conceptualization, Data curation, Formal analysis, Methodology, Visualization, Writing – original draft. **Dillard**: Conceptualization, Project administration, Supervision, Validation, Writing – Review & Editing.

1. Introduction

Resilience is an emergent property that conceptualizes a system's abilities to prepare, respond, recover, and adapt to disruptions in a way that does not compromise long-term functionality. As an organizing concept for planning and management, it has been applied under many contexts, from supply chains, which can span the globe, to communities, which provide a sense of place, belonging, or mutual support at the local scale. Community resilience can be especially difficult to measure because communities are complex systems with difficult to predict behavior. Consequently, indicators, which can synthesize the behavior of complex systems are increasingly used to measure and track community resilience. When properly implemented, community resilience indicators can provide evidence-based backing for planning and decision-making and also incorporate the priorities of local decision-makers, rights holders, and other interested parties [1]. Like many decision support tools, community resilience indicator frameworks rely on synthesis, which is broadly defined as an activity that searches for and interrelates data, information, and knowledge from multiple primary and secondary sources [2]. While standards and norms for generating various kinds of primary knowledge are relatively mature, guidance on synthesis is relatively less mature and fragmented across the many types of synthesis that exist across disciplines. Synthesis standards and norms are important because they help establish the credibility of knowledge production. Much progress has been made in some narrowly defined areas of synthesis, such as meta-analysis of randomized controlled experiments. However, in interdisciplinary fields, such as community resilience, it is often unclear where guidance should come from and how it should be combined.

Part of the challenge for community resilience indicator studies is that the definition and major factors driving community resilience are still open questions, leading to many studies combining two different activities: synthesis through conceptual modeling and measurement through indicator development. Consequently, many studies begin with constructing a logic model, which is a graphical or tabular description that depicts the interrelationships among relevant concepts. As flexible tools, logic models can incorporate causal interrelations and the broader context within which concepts exist, such as hierarchies or grouping [3, 4]. For community resilience indicators studies, indicators are linked to constructs, such as social capital and disaster recovery, that are interrelated by causation, hierarchy, and grouping. These constructs, which can initially be informed by background knowledge, are then used to search and review the relevant literature to refine the conceptual model and ultimately provide credibility for indicator inclusion/exclusion choices that appropriately summarize and measure construct behavior and metrics that quantify indicators.

Despite the frequent use of conceptual modeling for synthesis, community resilience indicator studies rarely cite or use guidance from relevant systematic review methods, such as framework synthesis, which is a widely-used method for building conceptual models from iterative literature review. An important concept in framework synthesis, and more broadly in systematic review, is that the studies included in a synthesis should be critically appraised for their methodological rigor and validity [5]. Also known as quality or weight of evidence, critically appraising evidence consists of a loose collection of techniques that vary widely in

approach [6]. At the simplest level, studies list out evidence from the literature and informally synthesize the degree to which the set supports a specific hypothesis or claim. The benefits of this style of appraisal are that it is relatively quick and can accommodate a large variety of hypotheses and study types, while its downsides are that the ultimate synthesis can be hard to reproduce and the judgments themselves may be subject to significant judgmental biases. More in-depth methods include rating schemes, such as Grading of Recommendations Assessment, Development, and Evaluation (GRADE) [7], and quantitative methods that attempt to model study quality and uncertainty, such as multi-criteria analysis [8]. While these methods can be more transparent and reproducible, their formality can easily be confused with rigor and their use can be more resource-intensive.

For the conceptual modeling of resilience constructs and indicators, community resilience studies most frequently follow the simplest option of informally synthesizing the literature with little use of critical appraisal. Similarly, the process of indicator development shows a lack of engagement with credibility-building practices typically found in measurement-theoretic fields, such as health and education [9]. Namely, these fields, which are well-versed in creating scales to measure unobserved constructs, use content validation to link indicator choices to the literature. Content validation is an iterative process that begins with creating an initial set of indicators through a literature search. These initial indicators are then presented to a set of suitable domain experts for further refinement through structured expert judgment. Depending on whether experts reach an agreement on indicator relevance (or lack thereof), the process may restart with further synthesis from the literature.

Content validation guidance emphasizes that this two-stage process—development and judgment—serves to balance the needs of creating a manageable set of indicators with validating that chosen indicators are indeed relevant and representative [10–12]. The content validation literature also cautions that utilizing only development or judgment can introduce threats to content and, ultimately construct, validity. Furthermore, if a study uses only one stage, then careful consideration should be given to how the chosen stage can provide a comprehensive but manageable set of indicators, as well as evidence that the set is also relevant. For community resilience indicators, which overwhelmingly rely on the development phase, this puts an increased burden on literature search and review to build credibility.

Synthesis and measure-theoretic indicator development offer promising and partially overlapping options for establishing credibility, but neither is solely capable of addressing the synthesis and measurement nature of community resilience indicator frameworks. This work's main contribution is to propose a rating scale that blends the two approaches, with a focus on efficiency, reproducibility, and rigor. To demonstrate its utility, a case study is also presented on a small selection of community resilience indicator frameworks and indicators and outline further avenues for expanding and applying the method.

2. Background

Part of the challenge of reconciling synthesis and measure-theoretic indicators development is that they typically address different areas of validity, which consists of construct, internal, and external (Fig. 1) [13].

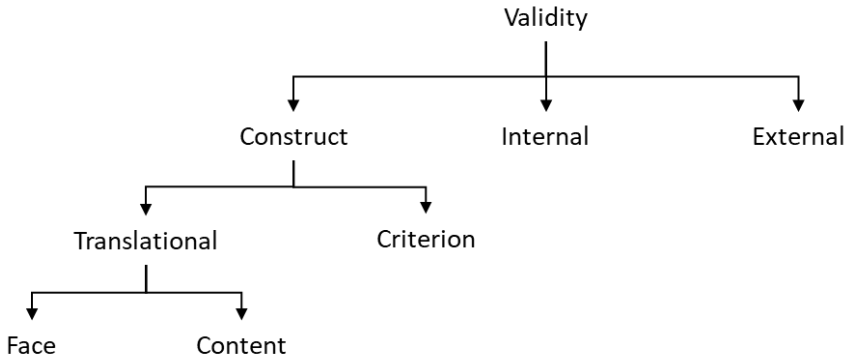

Fig. 1. Different types of validity.

Construct validity is a two-part concept that addresses whether a study measures the construct it claims to address. Measure-theoretic indicator development tends to focus on construct validity because indicators are focused on measuring unobservable constructs and properly establishing the indicators and constructs is vital to the credibility of the indicator set. The first part of construct validity, translational validity, provides evidence from the literature and/or expert judgment that variables have some link to underlying theory. It is composed of face and content validity. Face validity refers to whether indicators would appear appropriate to non-experts, while content validity establishes whether all relevant indicators have been included and all extraneous indicators have been excluded. The second part, criterion validity, includes familiar statistical tests, such as showing that an independent variable significantly correlates with the construct under question.

Whether or not an observed relationship has some rooting in theory is based on content validity, which is concerned with establishing the degree to which indicators of a framework or assessment are relevant to and representative of the construct a framework intends to measure [14]. In the judgment phase of content validity, the most common scale for judging relevance is the item content validity index (I-CVI) [15]. I-CVI is a four-point scale containing the descriptions: (1) not relevant, (2) somewhat relevant, (3) quite relevant, and (4) highly or very relevant. Variations of the I-CVI scale can contain three or five points if it is believed that having a neutral option is useful [10]. Another widely used scale, the content validity ratio (CVR) [16], contains three points: (1) not necessary, (2) useful but not necessary, and (3) essential.

In contrast, synthesis, which searches for and interrelates data, information, and knowledge from multiple primary and secondary sources, generally focuses on internal and external validity of the studies used. Internal validity assesses whether alternative explanations for an observed relationship have been ruled out and whether these determinations are subject to bias based on the type and specific implementation of a research method. External validity refers to the study's generalizability to a change in context, which includes, time period,

resolution, and scale; location and spatial resolution; experimental setup, and subject characteristics [17]. For example, using a study that has similar variables and constructs to a framework, but different locations (e.g., urban/rural) or subject characteristics (e.g., age, race), might introduce problems with external validity. Similarly, a study with similar locations and subject characteristics to a framework, but with different enough variables and constructs might also have external validity problems. Importantly, these judgments of external validity are independent of the study's construct or internal validity or their methodological rigor. A framework can use a study with high rigor and construct and internal validity, but an overall appraisal will suffer without external validity.

In this study, we argue that the concept of causality is a useful link between synthesis and measure-theoretic indicator development. To some extent, this is operationalized (but not explicit) in the use of 'necessary' in points (1) and (2) of CVR, which implies that an omitted 'essential' indicator would leave out a necessary causal relationship. In addition, establishing internal validity by ruling out alternative explanations is key to establishing causality [18, 19]. This link is useful because the synthesis literature has considerable depth in categorizing types of studies by their potential to imply causality and developing qualitative scales to the level of evidence for causality.

For example, the World Health Organization (WHO) International Agency for Research on Cancer (IARC) syntheses identify three streams of evidence—cancer in humans, cancer in experimental animals, and mechanistic—and pairs each with criteria for assigning qualitative descriptors [20, 21]. For example, sufficient evidence for cancer in humans involves observed covariance in which "chance, bias, and confounding were ruled out with reasonable confidence." In contrast, sufficient evidence for cancer in experimental animals is established when a causal relationship has been established in either two or more animal species, two or more independent studies of the same species, or in both sexes of the same species. Lines of evidence are integrated via an algorithm. For example, classifying an agent as carcinogenic in humans requires either a rating of sufficient for the cancer in humans stream or a combination of a limited or adequate rating of cancer in humans, a sufficient rating of cancer in experimental animals, and a strong rating for mechanistic evidence. Similarly, classifying an agent as probably carcinogenic can follow one of four combinations of evidence stream ratings.

Another example is from the U.S. Environmental Protection Agency's (EPA) National Ambient Air Quality Standards (NAAQS) syntheses, which use a six point scale—causal, likely to be causal, suggestive of, but not sufficient to infer causality, inadequate to infer causality, unlikely to be causal—to describe bodies of evidence for health, ecological, and other welfare effects [22–24]. Similar to IARC, NAAQS uses the degree of chance, confounding, or other biases to differentiate between ratings. However, it is different in that it does not have an algorithm to prescribe the number or type of studies. Rather, it is recommended that 'causal' and 'likely to be causal' ratings be determined by multiple high-quality studies.

While adopting an existing critical appraisal framework would be ideal, there are a few reasons why doing so might be challenging. For one, existing frameworks are designed for narrower, more well-defined problems than community resilience. In addition, existing frameworks also accept a narrower set of study types than might be used to justify the inclusion of resilience

indicators. Finally, a critical appraisal framework for resilience indicators serves a fundamentally different purpose. Many other frameworks are making final determinations that have significant regulatory consequences, and thus have significant resources for appraisal. Indicator frameworks only provide justification for including indicator measures, the performance of which should undergo criterion validation. Thus, there is a need to consider how the quality of evidence can be incorporated into a flexible and robust framework that is also resource-efficient.

3. Framework

3.1. Theoretical Underpinning

This study adopts a flexible approach rooted in Bayesian confirmation theory (BCT) that has a long history in using the mechanics of Bayes theorem to describe the informativeness of evidence [25]. In this application, evidence is provided by studies, such as peer-reviewed articles, that present links between observed variables (e.g., voting) and/or constructs (e.g., social capital or resilience) of interest. Framework researchers use evidence from the studies to infer a causal link between selected indicators and constructs. As studies are not always available that exactly match the indicators and constructs researchers are interested in, researchers use their judgment to determine how applicable any specific study is given similarities among variables and constructs, and study and framework context.

This study's implementation of BCT considers a Bayesian belief network of three discrete random variables—*H*, *S*, and *E*—all of which map the outcomes *false* and *true* to the real valued numbers 0 and 1. In the following description, lower case with a subscript of 0 or 1 indicates a false or true outcome, respectively. For example, h_1 is equivalent to $H(true) = 1$ and s_0 is equivalent to $S(false) = 0$. Similarly, the probability statements for *H* being *true*, $\mathbb{P}(H = 1)$ and $\mathbb{P}(h_1)$, are equivalent.

The random variable *H* represents the hypothesis that a specific framework indicator or construct is causally related to the construct of interest. The causal hypothesis considered in a specific study (*S*) is supported by evidence *E* in the study. Here, the outcome e_1 is consistent with the causal hypothesis being true. Relationships are assumed among the three variables such that *E* can be used to learn about *H* via *S* as long as *S* is not known with certainty—*H* and *E* are conditionally independent given *S*. This assumption clarifies that *E* is not the primary variable of interest, but rather whether *S* is true or not. The informativeness or diagnosticity of the study can be gauged by a ratio of conditional probabilities, also known as a Bayes factor (Ω), which is multiplied by the prior odds of causality ($O(h_1)$) to yield the posterior odds that an indicator or construct is causally related to resilience ($O(h_1|E)$):

$$O(h_1|E) = \frac{\mathbb{P}(E|S)\mathbb{P}(S|h_1)}{\mathbb{P}(E|S)\mathbb{P}(S|h_0)} O(h_1) = \Omega O(h_1) \quad (1)$$

In Eq. (1), the conditional probability distribution $\mathbb{P}(E|S)$ gauges the diagnosticity of the study itself, independent of its use for providing evidence for an indicator framework. It is specified by eliciting two conditional probabilities from Table 1, of which we believe the most intuitive are based on observing that a study reports a causal relationship (e_1). The first elicits the likelihood of observing confirming evidence assuming that the causal relationship tested by the study does actually exist, $\mathbb{P}(e_1|s_1) = \theta$, while the second elicits the likelihood of observing confirming evidence assuming the causal relationship tested by the study does not actually exist, $\mathbb{P}(e_1|s_0) = 1 - \eta$. In forming judgments about θ and $1 - \eta$, it can be useful to conceptually link these variables to traditional notions of statistical error, where θ is akin to the 'power' of the study to discover a causal link and $1 - \eta$ to the 'Type 1 error' of an erroneous result. However, when doing so, it is important to note that the judgments are more expansive

than traditional concepts of error due to randomness and include error due to biases and confoundedness.

Table 1. Probability table for evidence *E* conditional on a causal hypothesis *S*

	Tested causal relationship exists, s_1	Tested causal relationship does not exist, s_0
Study reports a causal relationship, e_1	θ	1 - η
Study does not report a causal relationship, e_0	1 - θ	η

The conditional probability distribution $\mathbb{P}(S|H)$ gauges the external validity of the study for providing evidence for the indicator framework. It is specified by eliciting two probabilities from Table 2, both of which are based on a thought experiment of assuming the causal relationship tested by the study does exist regardless of whether confirming evidence is reported. The first elicits the likelihood of the causal relationship being true assuming that the indicator is causally related to resilience $\mathbb{P}(s_1|h_1) = \lambda$. Conceptually, λ represents how likely the mechanisms and conditions that led to a causal relationship for the proposed indicator are the same as those that would lead to a causal relationship in the study. Differences in time period, place, and measures could all lead to $\lambda < 1$. The second elicits the likelihood of observing the causal relationship being true assuming that the indicator is not causally related to resilience, $P(s_1|h_0) = 1 - \gamma$. Conceptually, $1 - \gamma$ represents how likely the study causal relationship might exist for reasons independent of the mechanisms related to the indicator.

Table 2. Probability table for a causal hypothesis *S* conditional on an indicator hypothesis *H*

	Indicator has a causal relationship with resilience, h_1	Indicator does not have a causal relationship with resilience, h_0
Tested causal relationship exists, s_1	λ	1 - γ
Tested causal relationship does not exist, s_0	1 - λ	γ

With the relevant probabilities elicited, indicator framework researchers can calculate a study's informativeness through the resulting Bayes factor, depending on whether e_1 or e_0 is observed.

$$\Omega(h_1|e_1) = \frac{\theta\lambda + (1-\eta)(1-\lambda)}{\theta(1-\gamma) + (1-\eta)\gamma} \tag{2}$$

$$\Omega(h_1|e_0) = \frac{(1-\theta)\lambda + \eta(1-\lambda)}{(1-\theta)(1-\gamma) + \eta\gamma} \tag{3}$$

While indicator framework researchers could elicit precise judgments for the conditional probabilities and use Eq. (2) and Eq. (3) to report the resulting Bayes factor, we argue that this

simple setup is better suited as a semi-quantitative heuristic for describing evidence quality. First, it is common in expert judgment and synthesis methods to use qualitative descriptors to describe probability ranges, especially in cases where eliciting precise probabilities might lead to overconfidence. Figure 2, which is based on descriptors used in the Intergovernmental Panel on Climate Change (IPCC) process and reports [26], shows the set of descriptors recommended for the purpose of this study. This set captures a range of imprecise beliefs about likelihood. For example, if an expert is not confident enough to make an assessment for or against causality, then this can be communicated through *about as likely as not*, which encompasses both the precise sense of ignorance of 0.50 and the middle third of the probability distribution ($0.33 < p < 0.66$). If an expert has, at best, more confidence than even odds, then this can be expressed as *more likely than not* or *more unlikely than not*, depending on whether greater confidence is towards causality or not. Expressions of sharper probabilities tend to also imply more precise beliefs; as a result, probability ranges become more precise as probabilities sharpen to zero or one.

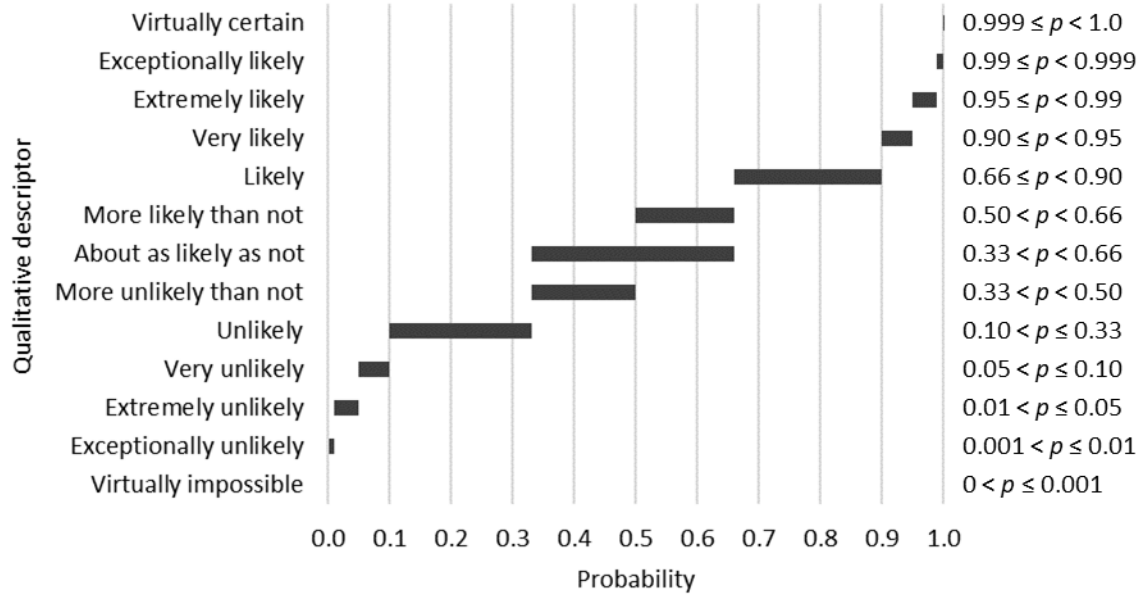

Fig. 2. Probability ranges of qualitative probability descriptors.

The second piece to a semi-quantitative framework are descriptors of Bayes factors and posterior odds. Table 3 reports a common scale and descriptors used for interpretation of Bayesian assessments that is based on Ref. [27]. As shown in Fig. 3, the intervals defined by the qualitative probability descriptors can be combined with Equation (2), to yield a mapping from beliefs about study quality to study informativeness. For example, a standard quantitative empirical study with no error due to confoundedness or other biases, would be described as being *likely* to be powerful enough to find a causal signal while being *extremely unlikely* to do so erroneously. If the study is *virtually certain* to be externally valid with respect to the indicator, then in observing a causal outcome in the study, one would conclude that the study provides *substantial* to *very strong* evidence that the indicator is causal. If more refinement is desired, then an author can consider whether their beliefs are towards the inner or outer boundaries of the qualitative probability interval. In this example, beliefs closer to the outer boundary, 0.01,

would yield *very strong* evidence, while those closer to the inner boundary, 0.05, would yield *substantial* evidence. Similar exercises may be less important for 'power' because the Bayes factor is much more sensitive to 'error'.

Table 3. Bayes factor ranges for qualitative evidence strength descriptors.

Bayes factor if e_1	Bayes factor if e_0	Qualitative descriptor
$D \geq 10^2$	$D \leq 10^{-2}$	Decisive
$10^{1.5} \leq D < 10^2$	$10^{-2} < D \leq 10^{-1.5}$	Very strong
$10^1 \leq D < 10^{1.5}$	$10^{-1.5} < D \leq 10^{-1}$	Substantial
$10^{0.5} \leq D < 10^1$	$10^{-1} < D \leq 10^{-0.5}$	Moderate
$10^0 \leq D < 10^{0.5}$	$10^{-0.5} < D \leq 10^0$	Marginal

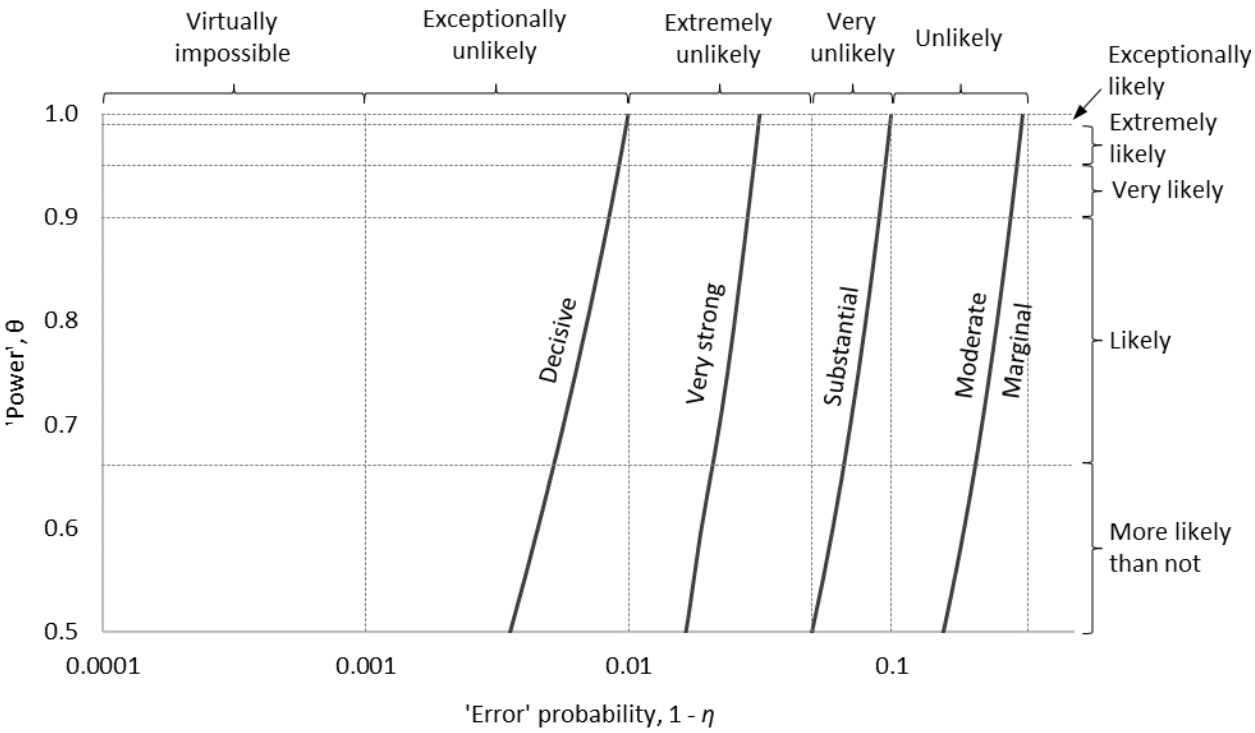

Fig. 3. Bayes factor for combinations of 'error' and 'power' when it is *virtually certain* (*virtually impossible*) an indicator is (not) causally related to resilience, given that a causal relationship exists. Solid lines mark boundaries between adjacent Bayes factor qualitative descriptors.

Figures 4-6 illustrate the sensitivity of the judgment to a study's external validity. For example, if a study is believed to be *exceptionally likely* to be externally valid, then the inner boundary ($\lambda = \gamma = 0.99$) of this interval results in no study—no matter how powerful or error-free—being able to provide *decisive* evidence (Fig. 4). At the inner boundary of the *extremely likely* external validity range ($\lambda = \gamma = 0.95$), no study is able to provide better than *substantive* evidence and a quantitative empirical study with perfect internal validity would provide

moderate to *substantive* evidence (Fig. 5). Furthermore, at the inner boundary of *very likely* external validity range ($\lambda = \gamma = 0.90$), no study can provide better than *moderate* evidence (Fig. 6), highlighting the importance of researchers carefully assessing the external validity of studies used to provide evidence. Similar to the differing sensitivity to 'power' and 'error' terms, the effect of the external validity parameters is largely driven by how likely the causal relationship in the study might exist for reasons independent of the mechanisms related to the indicator in the framework. Thus, researchers should put special attention on reasoning through the γ parameter.

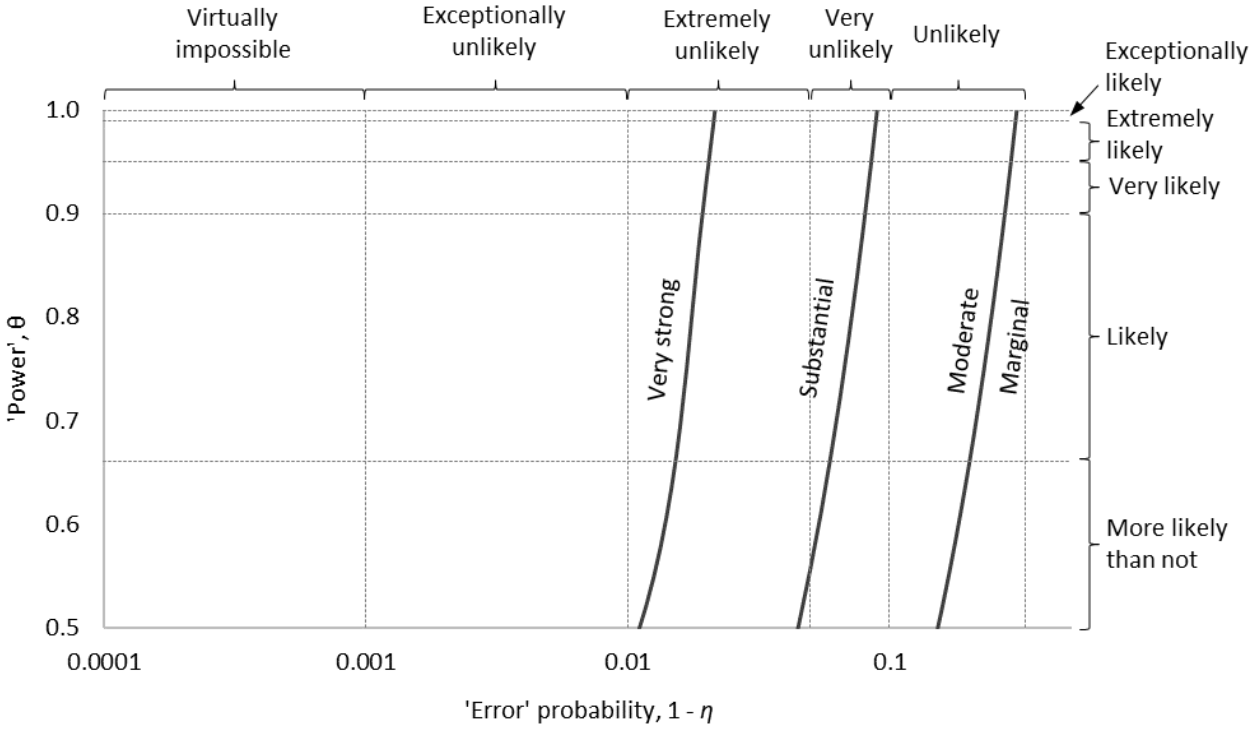

Fig. 4. Bayes factor for combinations of 'error' and 'power' when it is *exceptionally likely* (*exceptionally unlikely*) an indicator is (not) causally related to resilience, given that a causal relationship exists. Solid lines mark boundaries between adjacent Bayes factor qualitative descriptors.

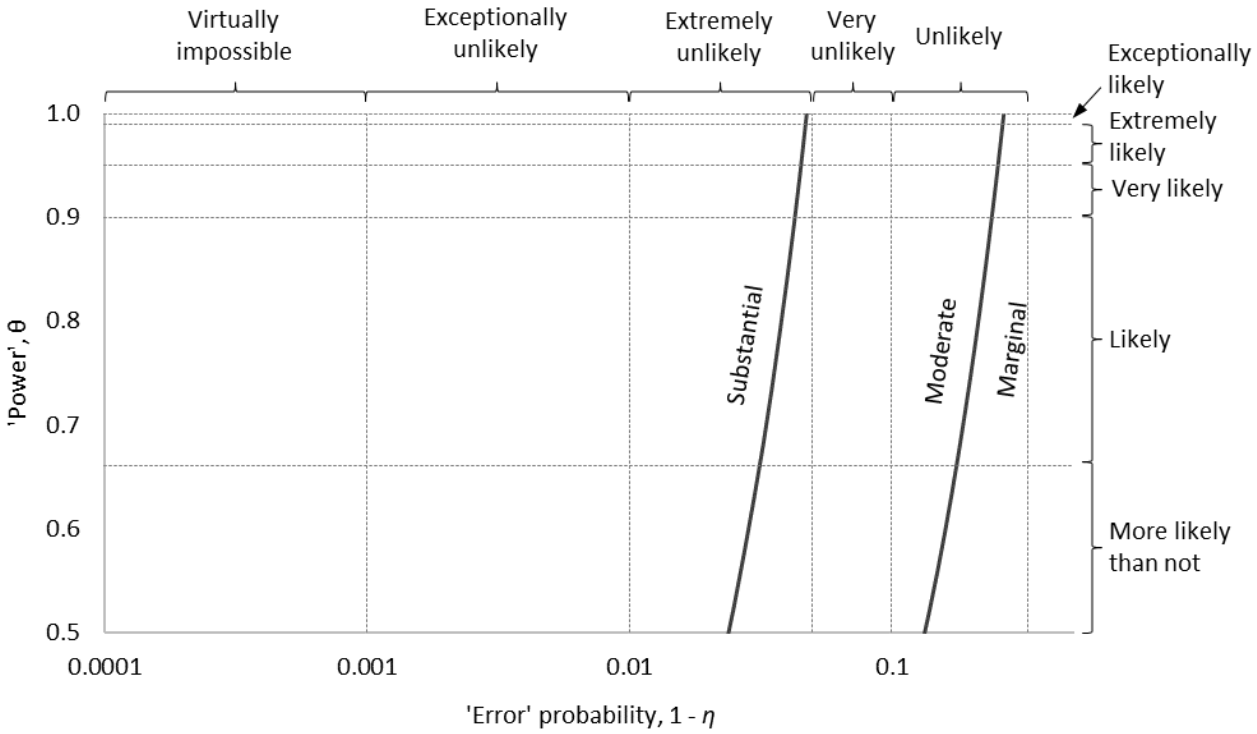

Fig. 5. Bayes factor for combinations of 'error' and 'power' when it is *extremely likely* (*extremely unlikely*) an indicator is (not) causally related to resilience, given that a causal relationship exists. Solid lines mark boundaries between adjacent Bayes factor qualitative descriptors.

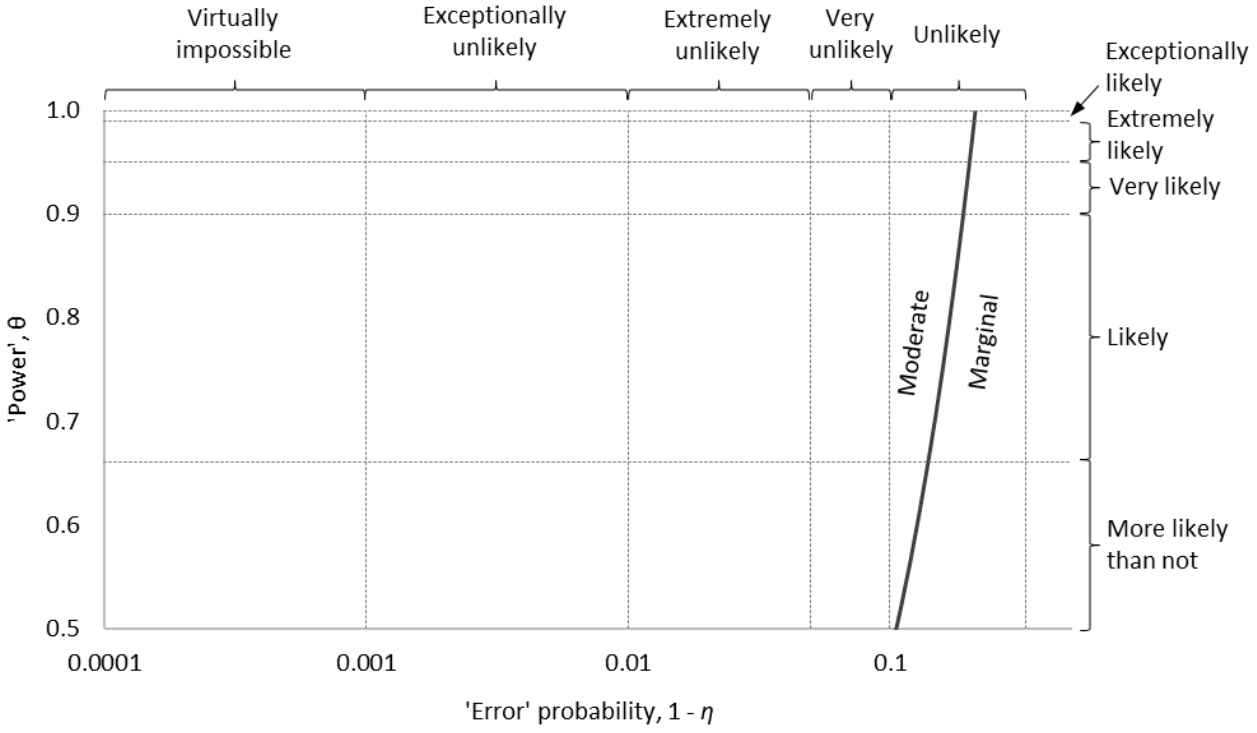

Fig. 6. Bayes factor for combinations of 'error' and 'power' when it is *very likely* (*very unlikely*) an indicator is (not) causally related to resilience, given that a causal relationship exists. Solid lines mark boundaries between adjacent Bayes factor qualitative descriptors.

3.2. Study Types

Unguided assessment of θ, η, λ, and γ has the potential to introduce judgmental biases. Fortunately, the causality, synthesis, and critical appraisal literatures have developed rough guides for how to categorize studies and consider the informativeness of their evidence [28–31]. In this study, a baseline level for each study type is utilized that can be adjusted upward or downward depending on the strength of the causality signal, violation of study assumptions, and external validity. This heuristic system is similar to other synthesis methods, such as GRADE [7], but differs in that our baseline and potential adjustments are directly linked to judgments related to the Bayes factor. Part of the advantage of this choice is that judgments are directly related to imprecise conditional probabilities, as opposed to qualitative grades that are difficult to interpret across different study types. In the following, the main study types are reviewed — quantitative empirical, qualitative empirical, mixed methods, and synthesis—and guidance is provided on their baseline levels and potentials for upgrading and downgrading the informativeness of the evidence.

3.2.1. Quantitative Empirical Evidence

Quantitative empirical studies use inductive methods designed to make general assertions from numeric data sampled from a population of interest. There are four types of inductive methods that vary depending on the assumptions required to establish causality (Table 4) [19, 32]. The typical design of quantitative empirical studies is such that they are *likely* to be powerful enough to detect a causal relationship and *extremely unlikely* to erroneously report a causality result due to randomness, confoundedness, and other biases. Consequently, their baseline rating is *substantial*, with the possibility of upgrading if a strong signal is observed and downgrading if external validity is less than *virtually certain*, standard unit treatment value assumption (SUTVA) is violated, or if other method-specific assumptions are violated or not discussed in the study. While the guidance presented does not prescribe specific amounts of downgrading, note that it does not take much 'error' or degraded external validity to lead to a lower rating. For example, even if a study is *virtually certain* of external validity, a *moderate* rating becomes possible if it is *very unlikely*, as opposed to *extremely unlikely*, that causality is erroneously reported (Fig. 2).

In the first type, which have the highest internal validity, investigators have control over the studies' experimental setting and sample units are selected at random or quasi-random. Experimental control means that confounding factors can be physically controlled for (e.g., in a laboratory), minimizing their likelihood to act as alternative explanations. Without randomization or quasi-randomization, the potential exists for the sample unit selection process to be correlated with independent and dependent variables, biasing the result.

Table 4. Types of quantitative empirical studies.

Study type	Baseline rating	Upgrade	Downgrade
Controlled experiments	Substantial	Strong signal (if e_1 is observed)	Poorly controlled or unconsidered confounding variables

			Lack of adjustment for threats to randomization Standard unit treatment value assumption is violated Less than *virtually certain* external validity
'As-if' randomized	Substantial	Strong signal (if e_1 is observed)	Exogenous assignment is irrelevant, not exclusionary, and/or nonmonotonic (natural experiments and instrumental variables) Continuity assignment is violated (regression discontinuity, interrupted time series) Standard unit treatment value assumption is violated Less than *virtually certain* external validity
Selection on unobservables	Substantial	Strong signal (if e_1 is observed)	Parallel trends assumption is violated (difference-in-differences) Standard unit treatment value assumption is violated Less than *virtually certain* external validity
Selection on observables	Substantial	Strong signal (if e_1 is observed)	Unconfoundedness assumption is violated Standard unit treatment value assumption is violated Less than *virtually certain* external validity

The empirical causality literature emphasizes that in absence of experimental control and/or randomization, there are three options for addressing internal validity [33, 34]. Among the types that provide the strongest evidence are those in which sample selection is controlled by a known exogenous process. If properly analyzed, such study types are said to be 'as-if' randomized and include natural experiments (NE), instrumental variables (IV), regression discontinuity (RD), and interrupted time series (ITS) methods [32, 35, 36]. Each method is subject to its own assumptions that should, if possible, be tested and discussed in the study. Natural experiments and instrumental variables methods share similar assumptions in that the exogenous assignment must be relevant, exclusionary, and monotonic. Regression continuity and interrupted time series share assumptions that discontinuity must be continuous [37, 38].

When these options are unavailable, so-called back and front door strategies are available [39]. Back-door strategies involve either (i) assuming assignment due to time-invariant, unobserved, confounding variables and controlling for these using difference-in-difference (DID) or fixed effects (FE) methods, or (ii) assuming all confounding variables are observable and controlling for these variables using statistical controls, matching, or stratification. Controlling for invariant unobserved variables is considered to require less fragile assumptions than assuming all confounding variables are observed (i.e., unconfoundedness assumption) and, therefore, is generally believed to provide better evidence for causality [37, 38, 40]. Front-door strategies involve identifying an exhaustive set of mechanisms that provide the only path between independent and dependent variables and are unrelated to unobserved variables that might affect both independent and dependent variables. The literature has noted that there are significant challenges in finding an exhaustive set of mechanisms that satisfy these assumptions and difficulties in showing that these assumptions hold. As a result, this strategy has not been extensively used [18, 41, 42].

3.2.2. Qualitative Empirical Evidence

The quantitative empirical studies discussed are based on post-positivist concepts of causality, in that it is assumed that causality is something that can be measured, however imperfectly, and that causality measured in similar contexts can, to some extent, be generalizable. The adoption of this philosophical stance can pose problems for using evidence from qualitative empirical research. While qualitative research is unified around using open-ended or subject-driven data collection methods to provide detailed data on humans' lived experiences, it includes a broad array of epistemological and ontological assumptions and analysis methods, many of which reject the notion that measuring generalizable causality is possible or desirable [43].

For the qualitative methods that do ascribe to a post-positivist stance on causality, it is generally agreed that establishing causality with the sole use of within or between case assessment leads to weak evidence for generalizing causal mechanisms beyond the case sample. This is because methods designed for between-case analysis excel at revealing broader patterns but can have difficulty ruling out confounding explanations. Inversely, methods designed for within-case analysis excel at offering detailed information in order to explore alternative hypotheses and rule out confounding explanations. However, as the within-case analysis applies only to a single case, generalizing across cases can be difficult. Thus, combining between-case and within-case methods balances out each methods' weaknesses in internal and external validity, respectively [44].

While there is not a consensus on how between-case and within-case methods should be combined, recent literature has begun to coalesce around set-theoretic multi-method approaches that investigate medium (12 to 50) to large (50+) number of cases [45]. These approaches are set-theoretic in that they focus on the relationships among collections of cases, as opposed to the average effects that are estimated in regression analysis. A prominent example is qualitative comparative analysis (QCA), which identifies combinations of mechanisms that lead to observed outcomes across cases, potentially delineating sufficient and/or necessary conditions. Discovered conditions are further investigated for causality by process tracing or similar methods on a judicious selection of cases from the sample [46]. Depending on the results, multi-method approaches may iterate with in-depth analysis of one or more cases or another QCA with adjusted assumptions.

Post-positivist qualitative methods have less developed guidelines or checklists for establishing causality than for quantitative methods. Regardless, there are recognized properties of higher quality set-theoretic multi-method approaches: clear rules for how cases were selected and developed, transparent choices behind QCA parameter choices, a strategy for which cases are selected for small-n within-case analysis, and sound reasoning for how within-case analysis feeds back into QCA results [47–51]. In addition, a large sample size of cases is not necessarily better in set-theoretic multi-method research, because case-based knowledge is needed to set up a QCA and select cases for within-case assessment [52, 53]. Some literature suggests that QCA robustness checks can mitigate a less in-depth knowledge of cases [54, 55], while others suggest that at best large-n QCA should be paired with quantitative methods in a mixed methods analysis [56].

Table 5. Types of qualitative empirical studies.

Study type	Baseline rating	Upgrade	Downgrade
Set-theoretic multi-method	Substantial	Strong signal (if e_1 is observed)	Too many conditions (large family wise error-rate) Too many robustness checks are failed Within-case analysis fails to eliminate alternative explanations Less than *virtually certain* external validity
Set-theoretic	Moderate	Strong signal (if e_1 is observed)	Too many conditions (large family wise error-rate) Too many robustness checks are failed Less than *virtually certain* external validity
Within case or small-n comparative	Moderate		Within-case analysis fails to eliminate alternative explanations Less than *virtually certain* external validity

For this study, it is considered possible for a qualitative study to provide the same baseline level of evidence as a quantitative study (Table 5). Specifically, set-theoretic multi-method studies have the potential to yield *substantial* evidence, but only if certain conditions are met; otherwise, the informativeness of the study should be downgraded. First, the family-wise error rate of QCA studies can quickly grow as the number of variables are considered. This is because each potential configuration of conditions is a separate hypothesis, much like interaction terms in regression analysis, and therefore studies should consider the likelihood of randomly arriving at results given the number of considered configurations [57, 58]. Second, QCA methods can be very sensitive to the cases included as well as selected algorithms and parameters. Robustness checks should be made to ensure that discovered configurations are not fragile to assumptions [54, 55]. Third, QCA on its own cannot provide information on causality; it must be paired with within-case analysis, such as process tracing. Studies should be downgraded if within-case analysis is not thorough or fails to eliminate confounding explanations through sufficiency and necessity tests [59]. All other qualitative methods are assumed to have a baseline rating of *moderate*. These include set-theoretic studies with no within-case analysis and within-case analysis only studies.

3.2.3. Mixed Methods Evidence

The literature has long acknowledged that, for some applications, quantitative or qualitative methods alone are not enough to properly understand both the breadth and nuances of a phenomenon. In these cases, various types of mixed methods studies have proven useful [44, 60]. For example, performing qualitative research before or at the same time as quantitative research can help illuminate issues that may lead to confounding and inform the appropriate quantitative design. In such a case, qualitative research can serve to improve the internal validity of the overall study by providing firmer evidence for statistical assumptions (e.g.,

informing the choice of instrumental variable), better understanding of heterogeneity across sample units, and potential measurement error in secondary data. Similarly, qualitative research conducted after quantitative analysis can bolster internal and construct validity by assessing how and why causal effects discovered in quantitative analysis occur.

A starting point for determining the strength of evidence for mixed methods studies is the quantitative or qualitative component with the highest rating. From there, questions can be asked as to whether the other component adds enough to the overall study to warrant upgrading the strength of evidence. For example, a mixed methods study with observed covariates analysis with a weak case for the unconfoundedness assumption as the quantitative component and a small-n comparative study as the qualitative component would initially be scored as *moderate*. Expert judgment would be used to decide if the qualitative component improves the case for the unconfoundedness assumption, possibly moving the mixed methods study overall rating to *substantial*. Similarly, expert judgment should be used to decide if two components at the same rating level warrant moving the overall rating up. For example, an observed covariates analysis with a strong case for the unconfoundedness assumption combined with a qualitative multi-method with medium or large-n might be complementary enough to move the overall study rating from *substantial* to *very strong*.

3.2.4. Synthesis Evidence

The synthesis evidence type encompasses a wide variety of studies (

Table 6). In reviewing the synthesis typology literature, two factors emerged that most strongly distinguish among synthesis types: synthesis logic and domain of generalizability [61–67]. Synthesis logic, which can be configurative or assimilative, refers to how individual pieces of evidence are related to one another. Configurative logic involves making inferences across mostly dissimilar pieces of evidence to create concepts and theories that best explain the data or identify gaps in the literature. In contrast, assimilative logic makes inductive inferences across similar pieces of evidence. Where configurative logic searches for coherence, assimilative logic confirms findings based on the repetition of comparable findings. The domain of generalizability describes where the synthesis is focusing its inference. In general, studies tend to either focus on a body of literature itself, highlighting gaps and trends or use the literature to generate theory or concepts or to test hypotheses.

Table 6. Types of synthesis studies.

Synthesis type	Synthesis subtype	Description	Considerations for assessing quality
Assimilative	Aggregative	Meta-analysis or -synthesis of similar studies	Standards of included study types Violation of aggregative synthesis type assumptions
	Cumulative	Exhaustive search of the literature to identify trends and gaps	Quality typically not assessed
Configurative	Interpretive	Non-exhaustive configurative synthesis	Quality typically not assessed
	Conceptual[a]	Indicator framework with only a conceptual model	Quality typically not assessed
	Suggested[a]	Indicator framework with suggested indicator metrics	Quality typically not assessed
	Quantified[a]	Indicator framework with quantified indicator metrics	Quality typically not assessed
	Validated[a]	Quantified indicator framework with criterion validation	Similar considerations to selection on observables
	Model[a]	Computational models that bring together data and information for exploration of phenomena or making predictions	Clearly states what purposes model is appropriate for Appropriate match between purpose and hypothesis Validation exercises that are appropriate for model and hypothesis Robustness checks
	All other[a]	Theory and framework development, perspectives	Quality typically not assessed
	Narrative	Non-exhaustive search of the literature in order to identify trends and gaps	Quality typically not assessed

[a]Indicates a subtype of interpretive synthesis

Assimilative synthesis tends to use well-documented exhaustive searches of the relevant literature and is more likely to assess quality of evidence. Within assimilative studies, there are two types: cumulative and aggregative. Cumulative synthesis is similar to narrative synthesis in that it provides generalized inferences about a group of studies or body of literature. It differs from narrative synthesis in that literature search is exhaustive and precise research questions are posed *a priori*. Evidence for these questions is amassed by assessing patterns in the relatively similar studies, where repeated patterns add credibility. Often this is done using various qualitative content analysis techniques, which can visualize the frequency and cross tabulation of themes and concepts. Aggregative synthesis pools data and findings to test a

specific hypothesis. In quantitative forms of aggregate synthesis, such as meta-analysis, formal statistical techniques pool data from primary studies to test hypotheses. Other forms, such as meta-synthesis, use only qualitative data or employ mixed-methods synthesis methods to incorporate quantitative and qualitative data.

In contrast, configurative synthesis tends to use representative, as opposed to exhaustive, coverage of the relevant literature and is less likely to assess quality of evidence. Within syntheses that use configurative logic, there are two types: narrative and interpretive. Narrative synthesis provides a comprehensive view on the current state of the literature, including gaps, but does not propose theory to explain what is found. Many classical literature reviews and background sections of primary studies are narrative syntheses in that they make general inferences about scientific gaps in the literature based on evidence collected through non-exhaustive literature search.

In this study, it is proposed that the best use of cumulative and narrative studies, which make inferences about the literature, are for sources of primary studies to use for evidence. The reason for this is that many synthesis studies do not properly assess the quality of evidence of included studies or synthesize them appropriately, potentially leading to a judgment of low evidence quality for a synthesis that may actually include high quality studies that provide informative evidence. For aggregative synthesis types, a baseline rating is proposed that is equivalent to included studies and adjusting upward or downward based on the strength and consistency of findings and violation or not of assumptions used in aggregative logic (e.g., statistical assumptions used in meta-analysis).

For interpretive synthesis, which configures diverse forms of evidence to produce new theory or build conceptual models, frameworks, or computational models, a more varied approach is proposed. For interpretive syntheses that recommend new theory, build conceptual models, construct guidelines, or offer a perspective, it is proposed using the frameworks as sources for studies as evidence, similar to cumulative and narrative syntheses.

An exception is that many indicator frameworks cite other frameworks as a basis for part or all of the included constructs and indicators. In this case, this methodology recommends categorizing indicator frameworks based on whether actual indicators are suggested, the indicators are quantified, and criterion validation is provided (

Table 6). For frameworks where criterion validation is provided, the same baseline and up/down-grading options should be used as for quantitative empirical studies that use selection on observable variables. For all others, the frameworks should be used as sources for studies, similar to cumulative and narrative synthesis studies, as opposed to considering the overall synthesis itself.

This framework also includes computational models under the category of interpretive synthesis, as these models use configurative logic to bring various pieces of information together to make predictions about phenomena that are difficult (e.g., future hazards in 50 years) or unethical to measure. Judging the quality of these types of syntheses varies widely depending on the model type and discipline. General checks are whether the study (i) fully examines the robustness of its findings by using alternative plausible model structures and parameters, (ii) clearly defines what the model is suitable for and appropriately matches model to hypotheses, and (iii) conducts validation exercises that are appropriate for the type of model and hypotheses.

4. Results

4.1. Test Case

The framework described in Section 3 is demonstrated by applying it to 18 common resilience indicators identified in Ref. [1], which reviewed seven implemented resilience indicator studies for common constructs and indicators (Table 7). All of the included frameworks are at the county-level, with all but two having coverage across the contiguous U.S. Four of the seven frameworks focus only on resilience, while one addresses both resilience and vulnerability (MRV) and another addresses wellbeing and resilience (NOAA). Notably, one framework, SoVI, uses vulnerability as its organizing construct instead of resilience. All but one of the frameworks (RCI) include at least 40 indicators; however, the range of indicators included varies widely, with CDRI at 97 and RCI at 17.

Table 7. Resilience indicator frameworks included in this study

Framework citation	Identifier	Framework description	Number of indicators in framework
Ref. [68]	BRIC	County-level resilience across the contiguous U.S.	49
Ref. [69]	CDRI	County-level resilience for the U.S. Gulf Coast	97
Ref. [70]	MitFLG	County-level resilience across the contiguous U.S.	42
Ref. [71]	MRV	County-level resilience and vulnerability across the contiguous U.S.	49
Ref. [72]	NOAA	County-level wellbeing and resilience for the U.S. Gulf Coast	64
Ref. [73]	RCI	County-level resilience across the contiguous U.S.	17
Ref. [74]	SoVI	Country-level vulnerability across the contiguous U.S.	42

The resilience indicator studies were coded by one researcher (Gerst), who first tagged citations used to justify indicator choices and then assessed the corresponding studies using our suggested framework. Coding decisions were checked by the other researcher (Dillard). Any disagreements were discussed and resolved until a consensus was reached.

4.2. Summary Statistics

The seven resilience indicator frameworks include a wide range of the most common indicators (Table 8). MRV includes the highest number of common indicators (15), while MitFLG includes the least (4). Six of the seven resilience indicator frameworks include at least half of the most common indicators. The mostly highly used indicators are not concentrated in one construct.

Educational attainment and healthcare labor force are both related to human capital, while inequality touches on economic systems and political engagement on institutional capital.

Table 8. Commonly used indicators and their inclusion in selected frameworks.

Indicator	Framework							Sum
	MRV	BRIC	SoVI	CDRI	NOAA	RCI	MitFLG	
Educational attainment	X	X	X	X	X	X		6
Healthcare labor force	X	X	X	X	X		X	6
Inequality	X	X	X	X	X	X		6
Political engagement	X	X	X	X	X	X		6
Civic organizations	X	X		X	X	X		5
Housing tenure	X	X	X	X		X		5
Regional economic vulnerability	X	X	X		X	X		5
Health insurance coverage	X	X		X		X		4
Healthcare infrastructure		X	X	X	X			4
Housing cost	X		X			X	X	4
Income			X	X	X		X	4
Labor force participation	X	X	X	X				4
Poverty	X		X		X	X		4
Unemployment	X		X		X		X	4
Age-based susceptibility	X	X	X					3
Telephone service		X		X	X			3
Vehicle access	X	X		X				3
Flood exposure	X							1
	15	13	13	12	11	9	4	

The resilience indicator frameworks exhibit different citation rates, as measured by the mean number of sources cited per indicator (Fig. 7). MitFLG, RCI, and MRV have on average less than one source of evidence per most commonly used indicator. In contrast, CDRI uses about two and NOAA uses over four sources of evidence per most commonly used indicator, indicating different strategies for presenting evidence for indicator choices.

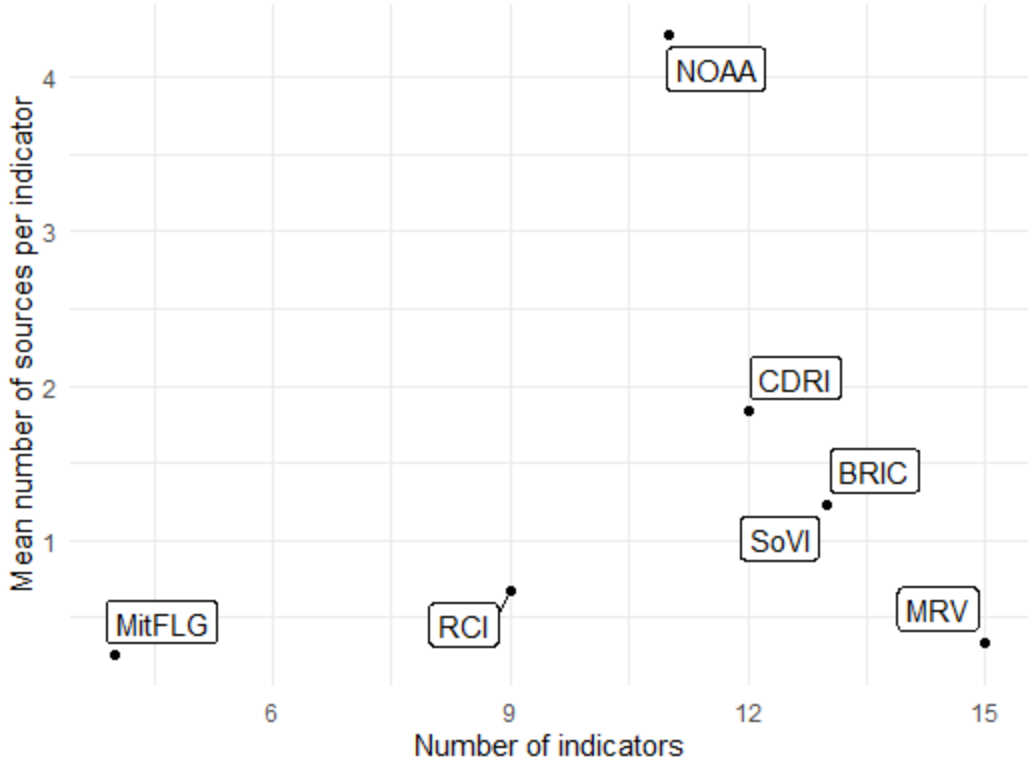

Fig. 7. Mean number of sources cited per indicator and number of indicators for each framework.

An application of this framework results in only 25 % of the citations passing on to the next stage of assessing evidence quality (validated indicator and empirical studies), as the other 75 % of citations are syntheses that are likely best used as sources for primary studies. Overall, the frameworks in this study base 84 % of their evidence on synthesis studies, with the balance (16 %) using empirical studies (Figure 8). At 29 % and 23 %, RCI and NOAA, respectively, utilize the highest fraction of empirical studies. MitFLG includes only syntheses, while CDRI includes just one empirical study. Within the citations to empirical studies, 94 % are quantitative, with the balance being mixed methods, and within the quantitative citations, 88 % are based on studies with selection on observable variables, with the balance being studies using as-if randomization.

In terms of synthesis logic type, configurative synthesis, which includes interpretive and narrative synthesis, comprise 96 % of the synthesis studies included in the frameworks, while the assimilative logic type, which includes aggregative and cumulative synthesis, comprises only 4 %. Furthermore, synthesis type is largely concentrated on interpretive synthesis, which comprises 90 % of overall synthesis studies. Within the interpretive synthesis type, 55 % of citations are to existing indicator frameworks. Citations vary across frameworks, with SoVI and CDRI drawing more from other interpretive synthesis types, such as studies that develop new theory or conceptual models, and the other frameworks relying heavily on existing indicator studies. The most common type of indicator study used are those that suggest indicators (36 %), followed by those that quantify indicators (32 %) and validate them (21 %). Use of these

types varies across the frameworks, with only four of the seven frameworks citing validated indicator studies.

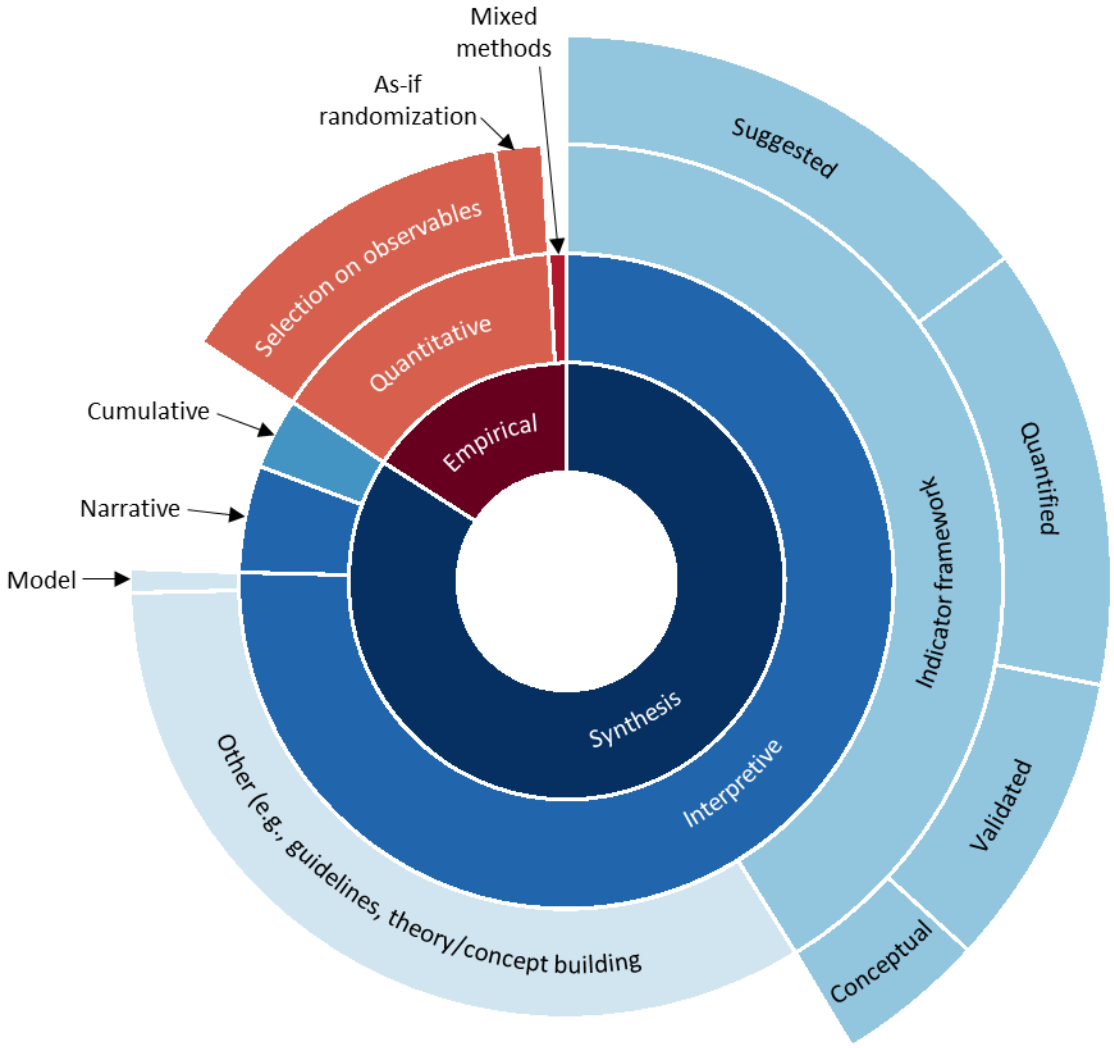

Fig. 8. Hierarchical representation of included study types. Moving from the center, division of each subsequent level is proportional to the adjacent type in the preceding level.

4.3. Assessing Validated Indicator Studies

Of the nine validated indicator studies cited, only four are related to disaster and resilience (Table 9). Three of these four (BRIC [68], CDRI [69], and SoVI [74]) are frameworks included in this study that are cited by other included frameworks, indicating a degree of circularity in the use of evidence. The other study, Sherreib et al. [75], is a foundational work on resilience indicators.

BRIC and Sherrieb both use convergent validation, which tests the correlation of the indicator index with another index that is thought to measure the same construct. Evidence quality of convergent validation is very dependent on the validity of the comparative index. In this case, BRIC and Sherrieb test their correlation with SoVI. The authors of SoVI use validation against a

concurrently occurring variable—the frequency of county-level disaster declarations over a decade—and find a weak correlational relationship, which infers weak convergent validation for BRIC and Sherrieb. In contrast, CDRI checks the concurrent relationship between its index and several measures of disaster impact, observing significant relationships. Consequently, BRIC, SoVI, and Sherrieb would likely be rated as having *marginal* evidence. If considering only the informativeness of CDRI, then it would be rated as *moderate*, as it has similar properties to a quantitative empirical study with selection on observables with potential confoundedness. However, the framework that cites CDRI does so for its transportation access and voter engagement indicators, the validity of which cannot be ascertained from the reported regressions in the CDRI study. As a result, the use of CDRI for this context, would also be rated as *marginal*.

The five validated indicator studies not related to disaster are cited to provide evidence for inclusion of: income and poverty [76–78], attachment to place [79, 80], education [79], and unemployment indicators [77, 78]. The studies by Osberg and Sharpe [77, 78] use potentially weaker divergent forms of validation, while Keyes [79] uses both convergent/divergent and predictive validation. The Osberg and Sharpe studies use divergent validation, which checks if an index does not correlate with an existing measure, to visually show that their alternative measure of economic well-being diverges from the established gross domestic product (GDP) measure. In contrast, the study by Keyes uses several established measures of social well-being to show convergent and divergent behavior, as well as survey answers on planned behavior to show predictive validation. Knight [80] uses a similar strategy for concurrent validation by using additional survey questions to correlate against its measured construct of attachment to place. Finally, Jia et al. [76] utilizes a more sophisticated multi-level model to concurrently validate proposed indicators of health-related quality of life. Resultantly, the Osberg and Sharpe studies would yield a *marginal* rating due to their relatively weak form of divergent validation, while the rest of the non-disaster indicator studies would be rated *moderate*.

Table 9. Application of framework to cited synthesis studies.

Study	Framework context	Study context	Validation type	Upgrade	Downgrade	Rating
Ref. [76]	Effect of income and poverty on economic security (NOAA)	Indicators of county-level health-related quality of life among U.S. adults	Concurrent validation with fixed effects to control causal heterogeneity	Strong signal	Selection on observable variables without assessment of confoundedness	Moderate
Ref. [79]	Effect of education on social well-being and attachment to place on social connectedness (NOAA)	Indicators of social well-being of U.S. adults	Convergent, divergent, and predictive validation	Convergent and divergent validation against established metrics	Selection on observable variables without assessment of confoundedness	Moderate
Ref. [80]	Effect of attachment to place on social capital as measured by tenure in community (NOAA)	Indicators of place attachment in 26 U.S. cities	Concurrent validation	–-	Simple correlational analysis	Moderate
Ref. [68] (BRIC)	Effect of primary care physicians on overall health and effect of unemployment on economic vitality (MitFLG) Measures of social resilience, evacuation capacity, vulnerable housing stock (MRV)	Indicators of county-level resilience in the U.S.	Convergent validation	–	Validation against weakly validated indicator framework (SoVI)	Marginal
Ref. [69] (CDRI)	Effect on resilience of transportation as measured by households with at least one vehicle and political engagement as measured by participation in presidential election (BRIC)	Indicators of county-level resilience in U.S. Gulf Coast communities	Concurrent validation		Mixed results of validation Lack of information to validate individual indicators used in index	Marginal

Table 9 continued. Application of framework to cited synthesis studies.

Study	Framework context	Study context	Validation type	Upgrade	Downgrade	Rating
Ref. [74] (SoVI)	Effect of reliance on natural resource sectors on economic vulnerability and measures of evacuation capacity (MRV)	Indicators of county-level social vulnerability in the U.S.	Concurrent validation	–	Weak correlation against concurrent damages	Marginal
Ref. [75]	Effect of education attainment equality, employment rate, racial income equality, independence from primary/tourism sectors, gender income equality, place attachment, political engagement, social capital, and population stability on resilience (BRIC)	Resilience indicators in Mississippi counties	Convergent validation	–	Weak signal	Marginal
Ref. [77,78]	Effect of poverty and unemployment on economic security (NOAA)	Indicators of economic well-being	Divergent validation	–	Weak form of divergent validation, relying on visual comparison	Marginal

4.4. Assessing Empirical Studies

From the seven frameworks reviewed for this study, 17 empirical studies were cited (Table 10). Of these, 14 involved selection using observable variables, two used as-if randomization, and one had mixed methods. While all these study types have a baseline rating of *substantial*, Refs. [81–91] were downgraded to *moderate* due to a combination of violated or unchecked causality assumptions and less than virtually certain external validity. None were rated above *substantial* and four received *marginal* ratings.

Two studies were rated *substantial*: one using as-if randomization and the other selection on observable variables. The RCI indicator framework used an as-if randomization setup by Ramcharan [92] to provide evidence for the effect of regional economic diversification on resilience as measured by the degree of difference from national economic diversification. Ramcharan used earthquakes as a natural experiment to select times and locations for estimating the effect of regional economic diversification on consumption shocks, along with two-way fixed effects to control for unobserved confounding. The study also included substantial robustness checks on its assumptions and found a strong and robust signal for the effect of diversification on mitigating shocks due to disaster. Based on the quality of the study, its use would have received a *very strong* rating, but in assessing its external validity there were questions about the generalizability of earthquake disasters to an indicator framework that applies to other disasters, such as flooding. Consequently, the quality of evidence for using Ramcharan in RCI is *substantial*.

Joynt [93] is the other study to receive a *substantial* rating. In this case, the NOAA indicator framework used Joynt to provide evidence for the effect of rurality in the U.S. Gulf Coast on availability and access to medical resources. Joynt studied the effect of rurality, among other variables, on quality of care for Medicare beneficiaries in the U.S., and while selection was based on observable variables, a strong signal was found, and a robustness check was conducted on the assumption of unconfoundedness. While the study was not perfectly externally valid, it was not so much so that the evidence rating should be downgraded.

The four studies that received *marginal* ratings are also instructive to highlight. The first, Brenner [94], which explores the effect of macroeconomic variables on health indicators in the mid-20th century U.S. It uses a minimally informative methodology in that it compares aggregate U.S. macroeconomic and health variables in a bivariate fashion with no formal statistical analysis or assessment of confoundedness. This method would likely receive a *marginal* rating. In addition, there are questions about its generalizability given that the indicator framework citing Brenner is applied to the U.S. Gulf Coast in the early 21st century.

The second study to receive a *marginal* rating is a mixed methods assessment of the effect of social isolation, among other variables, on elderly residents during the heat wave in Chicago in 1995 [95]. The qualitative component consisted of a small set of interviews and document analysis, which did not explore alternative explanations for many of the identified patterns. The quantitative analysis, which entailed visual comparison of maps, did not complement the lack

of confounding analysis in the qualitative component. Furthermore, it is unclear how generalizable the findings are to the NOAA framework context in the U.S. Gulf Coast, which has a very different climate and culture.

The third *marginal* study, Rupasingha [96], uses as-if randomization via instrumental variables and panel data to assess the effect of production factors (e.g., education) on U.S. county-level social capital measures. While it uses a study type with greater potential to establish causality, Rupasingha et al. does not test instrument assumptions and does not use fixed effects or differencing for its panel data, leading to questions on the as-if randomization assumption and whether it controls for unobserved confounding factors. If the study was generalizable to the indicator framework that cited it, the evidence would be *moderate*. However, the study that cited the framework, MRV, used Rupasingha et al. as evidence for the effect of social capital and civic engagement on social resilience in the U.S. as measured by number of 501(c)(3) organizations per capita and voter participation rate, which the study does not actually address. Thus, the overall rating was downgraded to *marginal*.

Tobin and Ollenberger [97] is the final study to receive a *marginal* rating, which the NOAA indicator framework used to provide evidence for the effect of employment status on post-disaster stress in U.S. Gulf Coast communities. Tobin and Ollenberger used selection on observable variables to assess the effect of demographic, social, and health factors on post-traumatic stress after flooding in Des Moines, IA. The study found significant but not strong evidence for the effect of unemployment and did not include any assessment of confounding variables. In addition, its external validity to coastal communities was less than virtually certain. In total, these three issues led to a downgrading to *marginal*. If the study had a much stronger signal, then a case could be made for a judgment of *moderate*.

Table 10. Application of framework to empirical studies.

Study	Framework context	Study context	Study type (baseline rating)	Upgrade	Downgrade	Rating
Ref. [92]	Effect of regional economic diversification on resilience as measured by the degree of difference from national economic diversification (RCI)	Effect of economic specialization on regional consumption shocks after an earthquake	'As-if' randomization using natural experiments (i.e., earthquakes) and two-way fixed effects (substantial)	Strong signal	Less than virtual certain of external validity	Substantial
Ref. [93]	Effect of rurality in the U.S. Gulf Coast on availability and access to medical resources (NOAA)	Effect of rurality on quality of care for Medicare beneficiaries in the U.S. in 2009	Selection on observable variables with hospital level fixed effects and confounder sensitivity analysis (substantial)	Strong signal	Less than virtually certain of external validity	Substantial
Ref. [81]	Access to medical care in U.S. Gulf Coast communities as measured by % of households without a vehicle (NOAA)	Effect of transportation access on health service utilization in rural western North Carolina in 1999	Selection on observable variables (substantial)	Strong signal	No assessment of confoundedness Less than virtually certain of external validity	Moderate
Ref. [82]	Effect of rurality in the U.S. Gulf Coast on availability and access to medical resources (NOAA)	Effect of rurality on availability and access to emergency care in the U.S. in the early 2000s	Selection on observable variables (substantial)	Strong signal	No assessment of confoundedness or measurement error of derived variables	Moderate

Table 10 continued. Application of framework to empirical studies.

Study	Framework context	Study context	Study type (baseline rating)	Upgrade	Downgrade	Rating
Ref. [83]	Investment in social capital due attachment to place in the U.S. Gulf Coast as measured by tenure in community (median year householder moved into unit) (NOAA)	Effect of tenure in community on attachment to place and well-being of over 50 residents in England in the early 2000s	Selection on observable variables (substantial)	Strong signal	No assessment of confoundedness Less than virtually certain of external validity	Moderate
Ref. [84]	Use of communication technology maintains and strengthens important social ties in the U.S. Gulf Coast (NOAA)	Effect of mobile phone and internet use on social network diversity and size in the U.S.	Selection on observable variables (substantial)	Strong signal	No assessment of confoundedness Study explicitly states that it is not measuring causality	Moderate
Ref. [85]	Economic vitality as a component of community economic resilience in the U.S. as measured by home ownership rates (% owner-occupied housing units) (BRIC)	Effect of home ownership on asset poverty in the U.S. in the 1980-90s	Selection on observable variables (substantial)	Strong signal	No assessment of confoundedness	Moderate
Ref. [86]	Effect of old age on lack of resilience in the U.S. (SoVI)	Effect of age on participation in emergency response following the Loma Prieta earthquake	Selection on observable variables (substantial)	Strong signal	No assessment of confoundedness Less than virtually certain of external validity	Moderate

Table 10 continued. Application of framework to empirical studies.

Study	Framework context	Study context	Study type (baseline rating)	Upgrade	Downgrade	Rating
Ref. [87]	Effect of social isolation on mental health in the U.S. Gulf Coast (NOAA)	Effect of social and civic activity on psychological well-being	Selection on observable variables (substantial)	–	No assessment of confoundedness	Moderate
Ref. [88]	Effect of regional affordability on economic security in the U.S. as measured by % of households paying less than 35% of income on housing (RCI)	Effect of personal vulnerabilities on precarious housing in the U.S.	Selection on observable variables (substantial)	Strong signal	No assessment of confoundedness Less than virtual certain of external validity	Moderate
Ref. [89]	Access to medical care in U.S. Gulf Coast communities as measured by % of households without a vehicle (NOAA)	Effect of race and other factors (e.g., income) on transportation difficulty for older residents of rural Alabama	Selection on observable variables (substantial)	Strong signal	No assessment of confoundedness	Moderate
Ref. [90]	Effect of social capital on well-being in U.S. Gulf Coast communities (NOAA)	Effect social capital on wellbeing in rural China	Selection on observable variables (substantial)	Strong signal	No assessment of confoundedness Explicitly correlational Less than virtually certain of external validity	Moderate
Ref. [91]	Effect of housing on resilience in U.S. Gulf Coast communities (CDRI)	Effect of damage and other factors on recovery of single-family housing in South Dade from Hurricane Andrew	Selection on observable variables (substantial)	Strong signal	Includes unobservable controls for time but not for the panel	Moderate

Table 10 continued. Application of framework to empirical studies.

Study	Framework context	Study context	Study type (baseline rating)	Upgrade	Downgrade	Rating
Ref. [94]	Effect of individual economic security in U.S. Gulf Coast on stress as measured by unemployment rate (NOAA)	Effect of macroeconomic variables in mid-20th century U.S., including unemployment, on health indicators	Selection on observable variables (substantial)	–	Simple bivariate timeseries correlational analysis with no assessment of confoundedness Less than virtually certain of external validity	Marginal
Ref. [95]	Effect of social isolation on risk of impacts from extreme events in the U.S. Gulf Coast (NOAA)	Effect of social isolation of elderly residents of Chicago during the 1995 heat wave	Mixed methods of interviews (qualitative) and visual comparison of spatial and temporal trends (quantitative)	–	Quantitative analysis uses visual comparison to highlight correlations Qualitative analysis does explore alternative explanations Less than virtually certain of external validity	Marginal
Ref. [96]	Effect of social capital and civic engagement on social resilience in the U.S. as measured by number of 501(c)(3) organizations per capita and voter participation rate (MRV)	Effect of production factors (e.g., education) on U.S. county-level social capital indices	'As-if' randomization using instrumental variables and panel data (substantial)	Strong signal	No assessment of endogeneity and instrument assumptions Panel data uses random effects *More likely than not* to hold external validity	Marginal
Ref. [97]	Effect of employment status on post-disaster stress in U.S. Gulf Coast communities (NOAA)	Effect of demographic, social, and health factors on post-traumatic stress after flooding in Des Moines, IA	Selection on observable variables (substantial)	–	No assessment of confoundedness Less than virtually certain of external validity	Marginal

5. Discussion

Application of this study's framework to the set of indicator frameworks from Ref. [1] yields ratings on primary and validated indicator studies that range from *marginal* to *substantial*. The lack of *very strong* and *decisive* levels occurs because of the prevalence of studies with unconfoundedness assumptions that are difficult to test and less than *virtually certain* external validity, both of which underscore challenges in finding evidence for indicator choices. First, data availability challenges in fields relevant to community resilience often lead to difficulties in performing controlled experimental, as-if randomized, or selection on unobservable variable studies. This leads to selection on observable variables being the most common option for indicator framework development. Second, community resilience indicators are place and scale (e.g., county versus city) specific, making it more difficult to find studies that reasonably match the spatial context of indicator frameworks.

Consequently, it is important for any evidence framework developed for community resilience indicators to avoid prescribing a minimum level of informativeness for a study to be included as evidence. Instead, all studies should be included and the indicator framework authors should appropriately communicate the quality of the evidence available to justify indicator choices, especially when further expert judgment will not be used for content validation. This study's framework provides one option for doing so that is based on researchers being able to identify study types, potential biases, and generalizability. This requires a workable understanding of the literature on causality and empirical study types. Within these requirements, this framework is not any more demanding than similar methods, such as GRADE.

One area where this study does not provide guidance is the minimum number of studies that should be used to justify any particular indicator choice or how to aggregate evidence from multiple studies. While there are Bayesian theoretic studies that address this question, the judgments required to estimate the conditional relationships among cited studies are likely too difficult for framework researchers to make [98–100]. As an alternative, framework researchers can 'triangulate' the evidence provided by a small set of studies. Triangulation, which is commonly used in epidemiology [101–103], involves comparing studies with different methods and contexts to draw stronger conclusions than could be drawn using any single study alone.

Triangulation works by comparing studies with different methods and, hence, different biases. If multiple studies with different biases all point towards causality, then this can provide stronger evidence for causality than any single study, especially if study biases have the potential to act in different directions. In addition, differences in study context can be used to triangulate the generalizability of findings. If no single study corresponds well with an indicator framework's geographic location and scale of measurement, then multiple studies that reach the same conclusion across different locations, time periods, and levels can improve the external validity of the gathered evidence. For the purpose of this proposed framework, researchers can implement triangulation by listing studies and their ratings along with a discussion of how study findings complement one another. In addition, the process used to find studies should be described, including to what degree the studies are representative of the broader literature, if a systematic search were to be conducted.

6. Conclusion

Part of the challenge for developing community resilience indicators is that the definition and major factors of community resilience are still open questions, leading to many studies combining two different activities: synthesis through conceptual modeling and measurement through indicator development. Implementing each activity's respective methods in one study is time-consuming, and as a result, indicator frameworks often neglect the full scope of assessing and communicating framework credibility.

Using Bayesian confirmation theory, this study offers a heuristic process for rating studies used to justify choices related to the conceptual model and indicators. By pre-defining a set of common study types, the framework does not require extensive expert judgment processes or systematic review, thus potentially reducing its burden. Demonstration of the framework was demonstrated for common indicators from a small set of existing resilience indicator studies. However, the efficacy of the framework will need to be further demonstrated by using it in practice as opposed to *ex post* examples.

The use of the framework in other domains should also be explored. The concept of assessing a study's evidence for causality is not limited to community resilience or creating indicators. Indicators are widely used in measuring complex systems, where direct measurement can be challenging. Thus, the potential use cases outside of community resilience are many, offering opportunities to refine the method. In addition, assessment processes, such as the National Climate Assessment [104], and other synthesis types, such as health interventions [105], all involve gauging the strength of a study's evidence. The heuristics outlined in this study may provide a helpful structure for these activities.

References

[1] Gerst MD, Dillard MK, Walpole EH, Loerzel J (2024) A review of community resilience indicators using a systems measurement framework. NIST Special Publication. (National Institute of Standards and Technology, Gaithersburg, MD), 2300–01. Available at 10.6028/NIST.SP.2300-01

[2] Tricco AC, Tetzlaff J, Moher D (2011) The art and science of knowledge synthesis. *Journal of Clinical Epidemiology* 64(1):11–20. https://doi.org/10.1016/j.jclinepi.2009.11.007

[3] Baxter S, Killoran A, Kelly MP, Goyder E (2010) Synthesizing diverse evidence: the use of primary qualitative data analysis methods and logic models in public health reviews. *Public Health* 124(2):99–106.

[4] Anderson LM, Petticrew M, Rehfuess E, Armstrong R, Ueffing E, Baker P, Francis D, Tugwell P (2011) Using logic models to capture complexity in systematic reviews. *Research Synthesis Methods* 2(1):33–42.

[5] Tod D, Booth A, Smith B (2022) Critical appraisal. *International Review of Sport and Exercise Psychology* 15(1):52–72.

[6] EFSA Scientific Committee, Hardy A, Benford D, Halldorsson T, Jeger MJ, Knutsen HK, More S, Naegeli H, Noteborn H, Ockleford C, Ricci A, Rychen G, Schlatter JR, Silano V, Solecki R, Turck D, Benfenati E, Chaudhry QM, Craig P, Frampton G, Greiner M, Hart A, Hogstrand C, Lambre C, Luttik R, Makowski D, Siani A, Wahlstroem H, Aguilera J, Dorne J, Fernandez Dumont A, Hempen M, Valtueña Martínez S, Martino L, Smeraldi C, Terron A, Georgiadis N, Younes M (2017) Guidance on the use of the weight of evidence approach in scientific assessments. *EFSA Journal* 15(8). https://doi.org/10.2903/j.efsa.2017.4971

[7] Guyatt GH, Oxman AD, Schünemann HJ, Tugwell P, Knottnerus A (2011) GRADE guidelines: A new series of articles in the Journal of Clinical Epidemiology. *Journal of Clinical Epidemiology* 64(4):380–382. https://doi.org/10.1016/j.jclinepi.2010.09.011

[8] Linkov I, Welle P, Loney D, Tkachuk A, Canis L, Kim JB, Bridges T (2011) Use of Multicriteria Decision Analysis to Support Weight of Evidence Evaluation. *Risk Analysis* 31(8):1211–1225. https://doi.org/10.1111/j.1539-6924.2011.01585.x

[9] Gerst MD, Dillard MK, Loerzel J (2025) Methodological recommendations for content validation of community resilience indicators. *Natural Hazards Review* 26(2). https://doi.org/10.1061/NHREFO.NHENG-2179

[10] Almanasreh E, Moles R, Chen TF (2019) Evaluation of methods used for estimating content validity. *Research in Social and Administrative Pharmacy* 15(2):214–221. https://doi.org/10.1016/j.sapharm.2018.03.066

[11] Lynn MR (1986) Determination and Quantification Of Content Validity: *Nursing Research* 35(6):382???386. https://doi.org/10.1097/00006199-198611000-00017

[12] Sireci S, Faulkner-Bond M (2014) Validity evidence based on test content. *Psicothema* (26.1):100–107. https://doi.org/10.7334/psicothema2013.256

[13] Broniatowski DA, Tucker C (2017) Assessing causal claims about complex engineered systems with quantitative data: internal, external, and construct validity. *Systems Engineering* 20(6):479–552.

[14] Haynes SN, Richard DCS, Kubany ES (1995) Content validity in psychological assessment: A functional approach to concepts and methods. *Psychological Assessment* 7(3):238–247. https://doi.org/10.1037/1040-3590.7.3.238

[15] Waltz CF, Bausell RB (1981) *Nursing research: design, statistics, and computer analysis* (F.A. Davis Co, Philadelphia).

[16] Lawshe CH (1975) A quantitative approach to content validity. *Personnel Psychology* 28(4):563–575. https://doi.org/10.1111/j.1744-6570.1975.tb01393.x

[17] Ferguson L (2004) External validity, generalizability, and knowledge utilization. *Journal of Nursing Scholarship* 36(1):16–22.

[18] Imbens GW (2020) Potential Outcome and Directed Acyclic Graph Approaches to Causality: Relevance for Empirical Practice in Economics. *Journal of Economic Literature* 58(4):1129–1179. https://doi.org/10.1257/jel.20191597

[19] Imbens GW (2024) Causal Inference in the Social Sciences. *Annual Review of Statistics and Its Application* 11(1):123–152. https://doi.org/10.1146/annurev-statistics-033121-114601

[20] Samet JM, Chiu WA, Cogliano V, Jinot J, Kriebel D, Lunn RM, Beland FA, Bero L, Browne P, Fritschi L, Kanno J, Lachenmeier DW, Lan Q, Lasfargues G, Le Curieux F, Peters S, Shubat P, Sone H, White MC, Williamson J, Yakubovskaya M, Siemiatycki J, White PA, Guyton KZ, Schubauer-Berigan MK, Hall AL, Grosse Y, Bouvard V, Benbrahim-Tallaa L, El Ghissassi F, Lauby-Secretan B, Armstrong B, Saracci R, Zavadil J, Straif K, Wild CP (2020) The IARC Monographs: Updated Procedures for Modern and Transparent Evidence Synthesis in Cancer Hazard Identification. *JNCI: Journal of the National Cancer Institute* 112(1):30–37. https://doi.org/10.1093/jnci/djz169

[21] IARC (2019) Preamble to the IARC Monographs on the Identification of Carcinogenic Hazards to Humans. (World Health Organization International Agency for Research on Cancer, Lyon, France), p 41.

[22] EPA (2015) Preamble to the Integrated Science Assessments (ISA). (U.S. Environmental Protection Agency, Washington, D.C.).

[23] Goodman JE, Prueitt RL, Harbison RD, Johnson GT (2020) Systematically evaluating and integrating evidence in National Ambient Air Quality Standards reviews. *Global Epidemiology* 2:100019. https://doi.org/10.1016/j.gloepi.2020.100019

[24] NASEM (2022) *Advancing the Framework for Assessing Causality of Health and Welfare Effects to Inform National Ambient Air Quality Standard Reviews* (National Academies Press, Washington, D.C.). https://doi.org/10.17226/26612

[25] Aitken C, Taroni F, Bozza S (2024) The Role of the Bayes Factor in the Evaluation of Evidence. *Annual Review of Statistics and Its Application* 11(1):203–226. https://doi.org/10.1146/annurev-statistics-040522-101020

[26] Mach KJ, Mastrandrea MD, Freeman PT, Field CB (2017) Unleashing expert judgment in assessment. *Global Environmental Change* 44:1–14. https://doi.org/10.1016/j.gloenvcha.2017.02.005

[27] Jeffreys H (1961) *Theory of Probability* (Oxford University Press, Oxford), 3d Ed.

[28] Pluye P, Gagnon M-P, Griffiths F, Johnson-Lafleur J (2009) A scoring system for appraising mixed methods research, and concomitantly appraising qualitative, quantitative and mixed methods primary studies in Mixed Studies Reviews. *International Journal of Nursing Studies* 46(4):529–546. https://doi.org/10.1016/j.ijnurstu.2009.01.009

[29] Pati D (2011) A Framework for Evaluating Evidence in Evidence-Based Design. *HERD: Health Environments Research & Design Journal* 4(3):50–71. https://doi.org/10.1177/193758671100400305

[30] Peterson MH, Barnason S, Donnelly B, Hill K, Miley H, Riggs L, Whiteman K (2014) Choosing the Best Evidence to Guide Clinical Practice: Application of AACN Levels of Evidence. *Critical Care Nurse* 34(2):58–68. https://doi.org/10.4037/ccn2014411

[31] Marquardt G, Motzek T (2013) How to Rate the Quality of a Research Paper: Introducing a Helpful Algorithm for Architects and Designers. *HERD: Health Environments Research & Design Journal* 6(2):119–127. https://doi.org/10.1177/193758671300600210

[32] Reeves BC, Wells GA, Waddington H (2017) Quasi-experimental study designs series—paper 5: a checklist for classifying studies evaluating the effects on health interventions—a taxonomy without labels. *Journal of Clinical Epidemiology* 89:30–42. https://doi.org/10.1016/j.jclinepi.2017.02.016

[33] Ferraro PJ, Sanchirico JN, Smith MD (2019) Causal inference in coupled human and natural systems. *Proceedings of the National Academy of Sciences* 116(12):5311–5318. https://doi.org/10.1073/pnas.1805563115

[34] Ferraro PJ, Hanauer MM (2014) Advances in Measuring the Environmental and Social Impacts of Environmental Programs. *Annual Review of Environment and Resources* 39(1):495–517. https://doi.org/10.1146/annurev-environ-101813-013230

[35] Dunning T (2008) Improving Causal Inference: Strengths and Limitations of Natural Experiments. *Political Research Quarterly* 61(2):282–293. https://doi.org/10.1177/1065912907306470

[36] Rockers PC, Røttingen J-A, Shemilt I, Tugwell P, Bärnighausen T (2015) Inclusion of quasi-experimental studies in systematic reviews of health systems research. *Health Policy* 119(4):511–521. https://doi.org/10.1016/j.healthpol.2014.10.006

[37] Bärnighausen T, Oldenburg C, Tugwell P, Bommer C, Ebert C, Barreto M, Djimeu E, Haber N, Waddington H, Rockers P, Sianesi B, Bor J, Fink G, Valentine J, Tanner J, Stanley T, Sierra E, Tchetgen ET, Atun R, Vollmer S (2017) Quasi-experimental study designs series—paper 7: assessing the assumptions. *Journal of Clinical Epidemiology* 89:53–66. https://doi.org/10.1016/j.jclinepi.2017.02.017

[38] Waddington H, Aloe AM, Becker BJ, Djimeu EW, Hombrados JG, Tugwell P, Wells G, Reeves B (2017) Quasi-experimental study designs series—paper 6: risk of bias assessment. *Journal of Clinical Epidemiology* 89:43–52. https://doi.org/10.1016/j.jclinepi.2017.02.015

[39] Pearl J (1995) Causal diagrams for empirical research. *Biometrika* 82(4):669–688. https://doi.org/10.1093/biomet/82.4.669

[40] Waddington H, White H, Snilstveit B, Hombrados JG, Vojtkova M, Davies P, Bhavsar A, Eyers J, Koehlmoos TP, Petticrew M, Valentine JC, Tugwell P (2012) How to do a good systematic review of effects in international development: a tool kit. *Journal of Development Effectiveness* 4(3):359–387. https://doi.org/10.1080/19439342.2012.711765

[41] Bellemare MF, Bloem JR, Wexler N (2024) The Paper of How: Estimating Treatment Effects Using the Front-Door Criterion*. *Oxford Bulletin of Economics and Statistics* 86(4):951–993. https://doi.org/10.1111/obes.12598

[42] Huber M (2024) An introduction to causal discovery. *Swiss Journal of Economics and Statistics* 160(1):14. https://doi.org/10.1186/s41937-024-00131-4

[43] Della Porta D, Keating M (2008) How many approaches in the social sciences? An epistemological introduction. *Approaches and Methodologies in the Social Sciences: A Pluralist Perspective* (Cambridge University Press, Cambridge, UK), pp 19–39.

[44] Harding, DJ, Seefeldt KS (2013) Mixed methods and causal analysis. *Handbook of Causal Analysis for Social Research*, Handbooks of Sociology and Social Research., ed Morgan SL (Springer, Dordrecht), pp 91–110.

[45] Beach D, Rohlfing I (2018) Integrating Cross-case Analyses and Process Tracing in Set-Theoretic Research: Strategies and Parameters of Debate. *Sociological Methods & Research* 47(1):3–36. https://doi.org/10.1177/0049124115613780

[46] Beach D (2020) Process Tracing Methods. *Handbuch Methoden der Politikwissenschaft*, eds Wagemann C, Goerres A, Siewert MB (Springer Fachmedien Wiesbaden, Wiesbaden), pp 699–719. https://doi.org/10.1007/978-3-658-16936-7_43

[47] Schneider CQ, Rohlfing I (2013) Combining QCA and Process Tracing in Set-Theoretic Multi-Method Research. *Sociological Methods & Research* 42(4):559–597. https://doi.org/10.1177/0049124113481341

[48] Schneider CQ (2023) *Set-Theoretic Multi-Method Research: A Guide to Combining QCA and Case Studies* (Cambridge University Press), 1st Ed. https://doi.org/10.1017/9781009307154

[49] Álamos-Concha P, Pattyn V, Rihoux B, Schalembier B, Beach D, Cambré B (2022) Conservative solutions for progress: on solution types when combining QCA with in-depth Process-Tracing. *Quality & Quantity* 56(4):1965–1997. https://doi.org/10.1007/s11135-021-01191-x

[50] Beach D (2018) Achieving Methodological Alignment When Combining QCA and Process tracing in Practice. *Sociological Methods & Research* 47(1):64–99. https://doi.org/10.1177/0049124117701475

[51] Daly J, Willis K, Small R, Green J, Welch N, Kealy M, Hughes E (2007) A hierarchy of evidence for assessing qualitative health research. *Journal of Clinical Epidemiology* 60(1):43–49. https://doi.org/10.1016/j.jclinepi.2006.03.014

[52] Greckhamer T, Misangyi VF, Fiss PC (2013) Chapter 3 The Two QCAs: From a Small-N to a Large-N Set Theoretic Approach. *Research in the Sociology of Organizations*, eds Fiss PC, Cambré B, Marx A (Emerald Group Publishing Limited), Vol. 38, pp 49–75. https://doi.org/10.1108/S0733-558X(2013)0000038007

[53] Finn V (2022) A qualitative assessment of QCA: method stretching in large-N studies and temporality. *Quality & Quantity* 56(5):3815–3830. https://doi.org/10.1007/s11135-021-01278-5

[54] Oana I-E, Schneider CQ (2024) A Robustness Test Protocol for Applied QCA: Theory and R Software Application. *Sociological Methods & Research* 53(1):57–88. https://doi.org/10.1177/00491241211036158

[55] Rutten R (2022) Applying and Assessing Large-N QCA: Causality and Robustness From a Critical Realist Perspective. *Sociological Methods & Research* 51(3):1211–1243. https://doi.org/10.1177/0049124120914955

[56] Fiss PC, Sharapov D, Cronqvist L (2013) Opposites attract? Opportunities and challenges for integrating large-n QCA and econometric analysis. *Political Research Quarterly* 66(1):191–198.

[57] Rohlfing I (2018) Power and False Negatives in Qualitative Comparative Analysis: Foundations, Simulation and Estimation for Empirical Studies. *Political Analysis* 26(1):72–89. https://doi.org/10.1017/pan.2017.30

[58] Braumoeller BF (2015) Guarding Against False Positives in Qualitative Comparative Analysis. *Political Analysis* 23(4):471–487. https://doi.org/10.1093/pan/mpv017

[59] Collier D (2011) Understanding Process Tracing. *PS: Political Science & Politics* 44(4):823–830. https://doi.org/10.1017/S1049096511001429

[60] Gillespie A, Glăveanu V, De Saint Laurent C (2024) *Pragmatism and Methodology: Doing Research That Matters with Mixed Methods* (Cambridge University Press), 1st Ed. https://doi.org/10.1017/9781009031066

[61] Mays N, Pope C, Popay J (2005) Systematically reviewing qualitative and quantitative evidence to inform management and policy-making in the health field. *Journal of Health Services Research & Policy* 10(1_suppl):6–20. https://doi.org/10.1258/1355819054308576

[62] Dixon-Woods M, Agarwal S, Jones D, Young B, Sutton A (2005) Synthesising qualitative and quantitative evidence: A review of possible methods. *Journal of Health Services Research & Policy* 10(1):45–53. https://doi.org/10.1177/135581960501000110

[63] Sandelowski M, Voils CI, Leeman J, Crandell JL (2012) Mapping the Mixed Methods–Mixed Research Synthesis Terrain. *Journal of Mixed Methods Research* 6(4):317–331. https://doi.org/10.1177/1558689811427913

[64] Gough D, Thomas J, Oliver S (2012) Clarifying differences between review designs and methods. *Systematic Reviews* 1(1):28. https://doi.org/10.1186/2046-4053-1-28

[65] Paré G, Trudel M-C, Jaana M, Kitsiou S (2015) Synthesizing information systems knowledge: A typology of literature reviews. *Information & Management* 52(2):183–199. https://doi.org/10.1016/j.im.2014.08.008

[66] Templier M, Paré G (2015) A Framework for Guiding and Evaluating Literature Reviews. *Communications of the Association for Information Systems* 37. https://doi.org/10.17705/1CAIS.03706

[67] Xiao Y, Watson M (2019) Guidance on Conducting a Systematic Literature Review. *Journal of Planning Education and Research* 39(1):93–112. https://doi.org/10.1177/0739456X17723971

[68] Cutter SL, Ash KD, Emrich CT (2014) The geographies of community disaster resilience. *Global Environmental Change* 29:65–77. https://doi.org/10.1016/j.gloenvcha.2014.08.005

[69] Peacock WG (2010) Advancing resilience of coastal localities: Developing, implementing, and sustaining the use of coastal resilience indicators: A final report. (Hazard Reduction and Recovery Center, College Station, TX), p 148.

[70] USDHS (2016) Mitigation Framework Leadership Group (MitFLG) Draft Concept Paper: Draft Interagency Concept for Community Resilience Indicators and National-Level Measures. (U.S. Department of Homeland Security, Washington, D.C.).

[71] Miller KK, Johnson A, Dabson B (2016) Measuring Resilience and Vulnerability in U.S. Counties. (University of Missouri Harry S. Truman School of Public Affairs Institute for Public Policy), Working Paper IPP/07.

[72] Dillard MK, Goedeke TL, Lovelace S, Orthmeyer A (2013) Monitoring Well-being and Changing Environmental Conditions in Coastal Communities: Development of an Assessment Method. (U.S. National Oceanic and Atmospheric Administration, Silver Spring, MD), NOAA Technical Memorandum NOS NCCOS 174, p 176.

[73] Foster K Resilience Capacity Index: Building Resilient Regions. (Institute of Governmental Studies: The University of California, Berkeley, Berkeley, CA).

[74] Cutter SL, Boruff BJ, Shirley WL (2003) Social Vulnerability to Environmental Hazards. *Social Science Quarterly* 84(2):242–261. https://doi.org/10.1111/1540-6237.8402002

[75] Sherrieb K, Norris FH, Galea S (2010) Measuring Capacities for Community Resilience. *Social Indicators Research* 99(2):227–247. https://doi.org/10.1007/s11205-010-9576-9

[76] Jia H, Moriarty DG, Kanarek N (2009) County-Level Social Environment Determinants of Health-Related Quality of Life Among US Adults: A Multilevel Analysis. *Journal of Community Health* 34(5):430–439. https://doi.org/10.1007/s10900-009-9173-5

[77] Osberg L, Sharpe A (2011) Moving from a GDP-based to a Well-being Based Metric of Economic Performance and Social Progress: Results from the Index of Economic Well-being for OECD Countries, 1980-2009. CSLS Research Report. (Centre for the Study of Living Standards, Ottawa), 2011–12.

[78] Osberg L, Sharpe A (2002) The index of economic well-being. *Indicators: The Journal of Social Health* 1(2):24–62.

[79] Keyes CLM (1998) Social well-being. *Social Psychology Quarterly* 61(2):121–140.

[80] Knight (2010) Soul of the Community: Why People Love Where They Live and Why it Matters: A National Perspective. (Knight Foundation, Miami, FL), p 36.

[81] Arcury TA, Preisser JS, Gesler WM, Powers JM (2005) Access to Transportation and Health Care Utilization in a Rural Region. *The Journal of Rural Health* 21(1):31–38. https://doi.org/10.1111/j.1748-0361.2005.tb00059.x

[82] Carr BG, Branas CC, Metlay JP, Sullivan AF, Camargo CA (2009) Access to Emergency Care in the United States. *Annals of Emergency Medicine* 54(2):261–269. https://doi.org/10.1016/j.annemergmed.2008.11.016

[83] Gilleard C, Hyde M, Higgs P (2007) The Impact of Age, Place, Aging in Place, and Attachment to Place on the Well-Being of the Over 50s in England. *Research on Aging* 29(6):590–605. https://doi.org/10.1177/0164027507305730

[84] Hampton KN, Sessions LF, Her EJ, Rainie L (2009) Social Isolation and New Technology: How the internet and mobile phones impact Americans' social networks. Pew Internet and American Life Project. (Pew Research Center, Washington, D.C.).

[85] Haveman R, Wolff EN (2004) The concept and measurement of asset poverty: Levels, trends and composition for the U.S., 1983?2001. *The Journal of Economic Inequality* 2(2):145–169. https://doi.org/10.1007/s10888-004-4387-3

[86] O'Brien P, Mileti DS (1992) Citizen participation in emergency response following the Loma Prieta earthquake. *International Journal of Mass Emergencies and Disasters* 10(1):71–89.

[87] Oliver JE (2003) Mental Life and the Metropolis in Suburban America: The Psychological Correlates of Metropolitan Place Characteristics. *Urban Affairs Review* 39(2):228–253. https://doi.org/10.1177/1078087403254445

[88] Pendall R, Theodos B, Franks K (2012) Vulnerable people, precarious housing, and regional resilience: an exploratory analysis. *Housing Policy Debate* 22(2):271–296. https://doi.org/10.1080/10511482.2011.648208

[89] Sook Park N, Roff LL, Sun F, Parker MW, Klemmack DL, Sawyer P, Allman RM (2010) Transportation Difficulty of Black and White Rural Older Adults. *Journal of Applied Gerontology* 29(1):70–88. https://doi.org/10.1177/0733464809335597

[90] Yip W, Subramanian SV, Mitchell AD, Lee DTS, Wang J, Kawachi I (2007) Does social capital enhance health and well-being? Evidence from rural China. *Social Science & Medicine* 64(1):35–49. https://doi.org/10.1016/j.socscimed.2006.08.027

[91] Zhang Y, Peacock WG (2009) Planning for Housing Recovery? Lessons Learned From Hurricane Andrew. *Journal of the American Planning Association* 76(1):5–24. https://doi.org/10.1080/01944360903294556

[92] Ramcharan R (2005) How big are the benfits of economic diversification? Evidence from earthquakes. IMF Working Paper. (International Monetary Fund, Washington, D.C.), WP/05/48, p 33.

[93] Joynt KE, Harris Y, Orav EJ, Jha AK (2011) Quality of Care and Patient Outcomes in Critical Access Rural Hospitals. *JAMA* 306(1). https://doi.org/10.1001/jama.2011.902

[94] Brenner MH (1977) Personal stability and economic security. *Social Policy* 8(1):2–4.

[95] Klinenberg E (1999) Denaturalizing disaster: A social autopsy of the 1995 Chicago heat wave. *Theory and Society* 28(2):239–295.

[96] Rupasingha A, Goetz SJ, Freshwater D (2006) The production of social capital in US counties. *The Journal of Socio-Economics* 35(1):83–101. https://doi.org/10.1016/j.socec.2005.11.001

[97] Tobin GA, Ollenburger JC (1996) Predicting levels of postdisaster stress in adults following the 1993 floods in the upper Midwest. *Environment and Behavior* 28(3):340–357.

[98] Landes J (2021) The variety of evidence thesis and its independence of degrees of independence. *Synthese* 198(11):10611–10641. https://doi.org/10.1007/s11229-020-02738-5

[99] Claveau F, Grenier O (2019) The variety-of-evidence thesis: a Bayesian exploration of its surprising failures. *Synthese* 196(8):3001–3028. https://doi.org/10.1007/s11229-017-1607-5

[100] Couch N (2022) The diversity principle and the evaluation of evidence. *Psychonomic Bulletin & Review* 29(4):1270–1294. https://doi.org/10.3758/s13423-022-02065-5

[101] Lawlor DA, Tilling K, Davey Smith G (2017) Triangulation in aetiological epidemiology. *International Journal of Epidemiology*:dyw314. https://doi.org/10.1093/ije/dyw314

[102] Pearce N, Vandenbroucke JP, Lawlor DA (2019) Causal Inference in Environmental Epidemiology: Old and New Approaches. *Epidemiology* 30(3):311–316. https://doi.org/10.1097/EDE.0000000000000987

[103] Thurmond VA (2001) The point of triangulation. *Journal of Nursing Scholarship* 33(3):253–258.

[104] Crimmins AR, Avery CW, Easterling DR, Kunkel KE, Stewart BC, Maycock TK eds. (2023) *Fifth National Climate Assessment* (U.S. Global Change Research Program, Washington, D.C.). https://doi.org/10.7930/NCA5.2023

[105] Pawson R, Greenhalgh T, Harvey G, Walshe K (2005) Realist review - a new method of systematic review designed for complex policy interventions. *Journal of Health Services Research & Policy* 10(1_suppl):21–34. https://doi.org/10.1258/1355819054308530