



Replications, Revisions, and Reanalyses: Managing Empirical Evidence in Software Engineering

Downloaded from: <https://research.chalmers.se>, 2026-10-26 00:44 UTC

Citation for the original published paper (version of record):

Frattoni, J., Fischbach, J., Fucci, D. et al (2026). Replications, Revisions, and Reanalyses: Managing Empirical Evidence in Software Engineering. FRONTIERS IN COMPUTER SCIENCE, 8.
<http://dx.doi.org/10.3389/fcomp.2026.1937190>

N.B. When citing this work, cite the original published paper.



OPEN ACCESS

EDITED BY

Mohammad Alshayeb,
King Fahd University of Petroleum and
Minerals, Saudi Arabia

REVIEWED BY

Shady Hegazy,
Siemens, Germany
Cathy Guevara-Vega,
Universidad Técnica del Norte, Ecuador

*CORRESPONDENCE

Julian Frattini
✉ julian.frattini@chalmers.se

RECEIVED 14 July 2026
REVISED 12 August 2026
ACCEPTED 28 August 2026
PUBLISHED 22 September 2026

CITATION

Frattini J, Fischbach J, Fucci D,
Unterkalmsteiner M, Mendez D, Feldt R
and Torkar R (2026) Replications,
Revisions, and Reanalyses: Managing
Empirical Evidence in Software
Engineering.
Front. Comput. Sci. 8:1937190.
doi: 10.3389/fcomp.2026.1937190

COPYRIGHT

© 2026 Frattini, Fischbach, Fucci,
Unterkalmsteiner, Mendez, Feldt
and Torkar. This is an open-access article
distributed under the terms of the
[Creative Commons Attribution License
\(CC BY\)](https://creativecommons.org/licenses/by/4.0/). The use, distribution or
reproduction in other forums is
permitted, provided the original author(s)
and the copyright owner(s) are credited
and that the original publication in this
journal is cited, in accordance with
accepted academic practice. No use,
distribution or reproduction is permitted
which does not comply with these terms.

Replications, Revisions, and Reanalyses: Managing Empirical Evidence in Software Engineering

Julian Frattini^{1*}, Jannik Fischbach^{2,3}, Davide Fucci⁴,
Michael Unterkalmsteiner⁴, Daniel Mendez^{3,4}, Robert Feldt^{1,5} and
Richard Torkar^{1,6}

¹Department of Computer Science and Engineering, Chalmers University of Technology and University of Gothenburg, Gothenburg, Sweden, ²Netlight Consulting GmbH, Munich, Germany, ³Fortiss GmbH, Munich, Germany, ⁴Software Engineering Research Lab, Blekinge Institute of Technology, Karlskrona, Sweden, ⁵Department of Communication, Quality Management and Information Systems, Mid Sweden University, Östersund, Sweden, ⁶Stellenbosch Institute for Advanced Study (STIAS), Stellenbosch, South Africa

Background: One aspired outcome of empirical research on quantitative data is a *variance theory*, i.e., a quantification of the effect of an independent on a dependent variables. In software engineering (SE) research, variance theories quantify—among others—the impact of tools, techniques, and other treatments on software development outcomes like productivity, cost-efficiency, and defect-detection rates. The validity of variance theories stems from the *synthesis* of multiple pieces of evidence, which increases its validity beyond the findings of a single study.

Gap: However, research synthesis in SE is rare and—if done—mostly limited to purely narrative syntheses. At best, researchers perform meta-analyses to synthesize variance theories from several quantitative results. But even meta-analyses only produce reliable results when synthesizing exact replications yet fail to generalize from variations.

Goal: We aim to extend the frontier of research synthesis beyond the state-of-the-art to systematically manage empirical evidence and its evolution.

Method: We apply method engineering to construct a framework for research synthesis from proven, individual method fragments. The framework allows researchers to put new evidence in a clear relation to an existing body of evidence and systematically expand knowledge about a studied phenomenon. We demonstrate the application of this framework to two fields of research by explicitly modeling the relationship between existing pieces of evidence.

Result: The framework puts three types of evolution of evidence into relation: (1) *replications* investigate the same hypothesis in a new context to improve external validity, (2) *revisions* challenge an existing hypothesis to improve internal validity, and (3) *reanalyses* replace analysis methods to improve conclusion validity. Through a systematic evolution of evidence and clear assessment criteria for each dimension of validity, the proposed framework can determine the frontier of a field of research.

Conclusion: The framework provides a perspective to systematically evolve empirical evidence in SE, supporting more constructive and productive advances in our field.

KEYWORDS

causal inference, replication, research synthesis, theory evolution, variance theory

1 Introduction

Software engineering (SE) research ultimately aims to benefit and support SE practitioners (Kitchenham et al., 2004; Budgen et al., 2013), partially by proposing and evaluating new tools, techniques, and approaches. This process of *knowledge translation* from research to practice requires both *knowledge creation* (i.e., gathering empirical evidence in primary studies) and *knowledge application* [i.e., providing this evidence to practitioners (dos Santos and Travassos, 2020)]. However, individual primary studies do not provide convincing decision support to practitioners on their own due to their limited scope and context (Lo et al., 2015; Devanbu et al., 2016; Franch et al., 2020; Juristo and Vegas, 2011). Hence, a necessary step between the creation and application of knowledge is the *synthesis* of primary studies.

One significant product of synthesizing quantitative research is a *variance theory*. Variance theories estimate the variance of a dependent variable caused by one or more independent variables (Ralph, 2018), i.e., the effect of independent on dependent variables. In SE research, variance theories provide decision support by quantifying the strength of the effect of, among others, tools, technologies, and human factors. For example, the synthesis of 27 primary studies about the effect of test-driven development (TDD) on code quality and developer productivity by Rafique and Mišić determined that “TDD has a small positive effect on quality but little to no discernible effect on productivity” (Rafique and Mišić, 2012).

However, research synthesis is rarely practiced in SE and, when done, mostly limited to narrative syntheses (dos Santos and Travassos, 2020). Only in very rare cases does SE research perform a *meta-analysis* which synthesizes multiple pieces of quantitative evidence (dos Santos and Travassos, 2020). Yet, even though certain forms of meta-analysis excel at synthesizing evidence from quantitative studies, they only produce usable results under certain conditions (Pickard et al., 1998). For example, if individual pieces of evidence come to contradictory conclusions, a meta-analysis would be inconclusive and the resulting variance theory could not offer decisive decision support (Hayes, 1999). In addition, current research synthesis practices struggle to accommodate more complex relationships between individual pieces of evidence: Interaction or mediation effects can be identified post-hoc only if candidate mediators are consistently operationalized and recorded (Rafique and Mišić, 2012; Yang, 2002), and many meta-analysis methods invite the unreflected usage of statistical techniques (dos Santos and Travassos, 2020; Kitchenham et al., 2020). As a consequence, resulting variance theories can be invalid and even detrimental when practitioners use them for their decisions.

In this work, we aim to extend the frontier of current research synthesis practices in SE beyond traditional meta-analyses. We apply method engineering to assemble a conceptual framework from several fragments of meta-research that address the current shortcomings in research synthesis practices (Section 4). We formally define quantitative, empirical evidence (Section 4.1), how two pieces of evidence can relate to each other (Section 4.2), and derive how variance theories emerge and evolve systematically

(Section 4.3). Finally, we demonstrate the framework by applying it to active fields of SE research (Section 5). After discussing our results (Section 6) we conclude (Section 7) with the aim to extend the perspective on research synthesis to obtain more valid variance theories from SE research. While a useful method for synthesizing existing evidence, the proposed framework holds particular value to guide researchers toward more coherent and productive research agendas in the future.

2 Related work

Research synthesis (Section 2.1) is a necessary step in empirical software engineering to produce variance theories (Section 2.2), a valid and reliable estimation of how variables affect other variables. Variance theories aim to have greater *validity* than single pieces of evidence yet often still fail to achieve it (Section 2.3) which motivates our approach.

2.1 Research synthesis

The purpose of any endeavor in SE research is to support SE practitioners (Budgen et al., 2013). This requires translating knowledge created in research to practice (Beecham et al., 2014; Cartaxo et al., 2016). However, previous research has shown that individual empirical studies do not provide convincing evidence to practitioners (Lo et al., 2015). A single study cannot compete with the beliefs of practitioners, as the findings are limited by its context (Devanbu et al., 2016). As Shepperd notes, “It would be highly unusual for a single primary study to provide a definitive answer to any question of much significance to practitioners” (Shepperd, 2013).

Consequently, many scholars advocate that “we need to combine results from multiple primary studies” (Shepperd, 2013) and that the field of SE research “needs to move to a portfolio of empirical studies (on a single research hypothesis) being the norm rather than the currently unconvincing ‘one-off’, normally laboratory-based, studies that currently dominate the research literature” (Miller, 2005) and still do (Shepperd et al., 2018). These portfolios of replications [when replicating a controlled experiment also called *families of experiments* (Basili et al., 1999)] strengthen the validity of research findings and generate more reliable conclusions (Nosek and Errington, 2020).

While portfolios of replications strengthen the knowledge creation phase of knowledge translation (dos Santos and Travassos, 2020) by increasing the validity of drawn conclusions, they do not necessarily serve the knowledge application phase. Indeed, portfolios of replications on their own even complicate maintaining an overview of all relevant primary studies, and contradicting results are non-trivial to reconcile (Rosenthal and DiMatteo, 2001). This necessitates the aggregation and integration of results from primary studies (Ciolkowski and Münch, 2005), commonly referred to as *research synthesis*. Research synthesis is a “collective term for a family of methods that are used to summarize, integrate, combine, and compare the findings” (Cruzes and Dybå, 2011)

of individual pieces of evidence about the same phenomenon. Shepperd summarized the synthesis process in five steps (Shepperd, 2013).

1. **Problem formulation:** specifying a research question, usually about the influence of one independent variable on one dependent variable.
2. **Locating evidence:** searching literature that contains evidence about the research question.
3. **Appraising evidence quality:** applying quality inclusion criteria.
4. **Evidence synthesis and interpretation:** extracting relevant data and performing the research synthesis that infers about the initial research question based on the accumulated evidence.
5. **Reporting:** disseminating the results in a research report.

To implement step four, SE research has adopted several research synthesis methods from more mature empirical research disciplines (Dixon-Woods et al., 2005; Pickard et al., 1998; Shepperd, 2013; Dieste et al., 2011; dos Santos and Travassos, 2020; Santos et al., 2019). Among the most popular is the *narrative synthesis* (Santos et al., 2018), a textual summary of research findings that is often criticized for its unsystematic approach (Ciolkowski, 2009) and lack of a quantitative outcome (i.e., a quantification of the effect strength of the observed phenomenon). The more systematic *vote-counting* classifies the effect of an independent variable on a dependent variable on ordinal scales, e.g., depending on its sign (i.e., positive, neutral, or negative effect) (Pickard et al., 1998) or its strength of evidence (e.g., third party claims, circumstantial evidence, and strong evidence) (Wohlin, 2013). A histogram of classified findings from primary studies indicates the tendency of the effect observed by a portfolio of replications. Similarly, the *structured synthesis method* (SSM) proposed by dos Santos and Travassos (2013, 2017) aggregates evidence from heterogeneous sources based on the belief value of these primary studies.

Empirical researchers usually consider *meta-analysis* as the state-of-the-art for synthesizing quantitative results from controlled experiments (Rosenthal and DiMatteo, 2001). Other than vote-counting or the SSM, meta-analysis solely integrates reported statistics without introducing any researcher subjectivity beyond the selection of primary studies. Various forms of meta-analysis exist and are adopted in SE research (Santos et al., 2018). The two approaches considered superior are *aggregate data* (AD) meta-analysis and *stratified independent participant data* (IPD-S) meta-analysis (Santos et al., 2018). AD meta-analysis pools the calculated effect sizes of primary studies together and calculates an overall effect size of the independent on the dependent variable (Rosenthal and DiMatteo, 2001), often represented in a Forest plot (Lewis and Clarke, 2001). IPD-S meta-analysis pools together the raw data from all experiments of the primary studies and analyzes this data directly (Riley et al., 2010), but retains information about the belonging of each data point to the experiment it originated from to account for between-study variance (Santos et al., 2018). IPD-S meta-analysis is commonly regarded as the gold standard for research synthesis (Sutton and Higgins, 2008) but is rarely applied in SE research (Santos et al., 2018) since it depends on the availability of the raw data from every included primary study. Among the

different methods, narrative synthesis and AD meta-analysis remain most present in SE research (Santos et al., 2018). Yet, meta-analysis remains uncommon in SE in general (Kitchenham et al., 2020). Some examples from our field of research include synthesizing evidence about defect prediction (Hosseini et al., 2017; Zain et al., 2023) and TDD (Rafique and Mišić, 2012).

2.2 Theory building

Postulating and validating theories is a goal of any scientific discipline (Fernández and Passoth, 2019) to abstract from passing trends and provide practitioners with reliable, valid decision support (Hannay et al., 2007). SE mainly considers three types of theories (Ralph, 2018):

1. **Theories for understanding** organize entities into meaningful categories.
2. **Process theories** explain how something is happening.
3. **Variance theories** quantify the strength of the effect of an independent variable on a dependent variable.

Each type of theory serves a distinct purpose. Variance theories quantify the effect strength of a phenomenon, which offers decision support to practitioners, e.g., when deciding whether to adopt a technique like TDD (Rafique and Mišić, 2012). To achieve this quantification, researchers need to synthesize individual pieces of quantitative evidence about a phenomenon.

2.3 Validity of variance theories

A SE practitioner could consult either a variance theory or a single piece of quantitative evidence for decision support. The appeal of a variance theory over a single piece of evidence is not just the summary of relevant knowledge, but also that a variance theory aims to be more *valid* than a single piece of evidence (Hannay et al., 2007). Because variance theories—just like controlled experiments—aim to quantify the effect strength of one or more independent variables on a dependent variable, the four dimensions of validity proposed by Cook et al. (1979) and popularized in SE by Wohlin et al. (2012) apply similarly. In the following subsections, we elaborate how variance theories aim to exceed individual studies in each dimension of validity. We close each subsection with current limitations that motivate advancing the frontier of research synthesis.

2.3.1 External validity

Most notably, a variance theory exceeds a single piece of evidence in *external validity*. External validity describes the degree to which a drawn conclusion can be generalized from the studied to the larger, often industrial context (Wohlin et al., 2012). Since a variance theory synthesizes evidence from more than one study, it can cover a context larger than any individual study. As such, the effect concluded by a variance theory is more likely to generalize to the target population (Gómez et al., 2014).

However, producing replications eligible for synthesis is non-trivial as their configuration can vary across several dimensions, e.g., operationalization, population, protocol, and experimenters (Gómez et al., 2014). This creates a spectrum ranging from replications as close as possible to the original study [referred to as literal (Gómez et al., 2014) or strict replications (Juristo and Gómez, 2012)] over replications varying one dimension [referred to as operational (Gómez et al., 2014) or differentiated replications (Juristo and Gómez, 2012)], to replications varying all dimensions and only keeping the underlying hypothesis, i.e., conceptual replications (Gómez et al., 2014; Juristo and Gómez, 2012). How the degree of variation between studies strengthens or threatens the external validity of a synthesized variance theory has been subject to extensive debate (Kitchenham, 2008; Juristo and Vegas, 2011). Gomez et al. consolidated the debate with the proposal of *systematic replications*, which structure multiple replications varying different dimensions to purposefully address potentially influencing factors (Gómez et al., 2014). But even these systematic replications could only detect *if* any dimension influences the phenomenon of interest but not *how*. In such a case, a systematic replication would have identified a relevant effect but remains unable to quantify its significance.

2.3.2 Internal validity

A variance theory can also be more *internally valid* than single pieces of evidence. Internal validity describes the degree to which the drawn conclusion is causal and free of confounding effects (Wohlin et al., 2012). Researchers usually achieve this by revising an existing hypothesis via adding or changing its causal assumptions. For example, in the field of software testing, Inozemtseva and Reid revised the common claim that the coverage of a test suite correlates with its effectiveness by postulating that the *test suite size* confounds this effect (Inozemtseva and Holmes, 2014).

This populates the literature about a specific phenomenon with evidence of varying internal validity. In principle, research synthesis can maximize the internal validity of a variance theory by constraining the evidence location phase (cf. Section 2.1) to only those pieces of evidence with maximum internal validity. However, current research synthesis methods are not fit to act on these variations. This is rooted in their focus on relationships between exactly two variables (Miller, 2000; Santos et al., 2018). Yet, it is unrealistic to assume that contemporary phenomena are isolated to only two variables, as human (Pickard et al., 1998) and other context factors (Pfleeger, 1999) usually interact with a phenomenon of interest.

Attempts to bend meta-analysis to deal with more complex causal assumptions beyond two-variable relationships produced subgroup or stratified analyses. For example, Harris and Rosenthal (1985) clustered a synthesis study with 136 primary studies into 31 meta-analyses to isolate the effect of mediators. Similarly, the aforementioned meta-analysis by Rafique and Mišić (2012) involving 27 primary studies allowed a subgroup analysis revealing the mediating effect of the study population. However, all of these approaches are post-hoc attempts that insufficiently compensate

the lack of an upfront, causal understanding of complex hypotheses. Despite their additional level of complexity, they still fail to capture moderating or interaction effects (Frattini et al., 2026).

2.3.3 Conclusion validity

Variance theories can also improve the *conclusion validity* of a claim over a single piece of evidence. Conclusion validity describes the degree to which statistical assumptions are met when applying inferential statistics to derive a quantitative conclusion from data (Wohlin et al., 2012). When the raw collected data is available—e.g., in IDP-S meta-analyses—this data can be analyzed with the best-fitting statistical method, regardless of what method has been applied to the individual pieces of evidence. Such improvements to conclusion validity are typically driven by adopting more robust statistical methods (McElreath, 2018), a method often referred to as a *reanalysis* or *test for robustness* (Nosek and Errington, 2020). For example, Furia et al. (2019) reanalyzed two empirical studies about the effectiveness of automatically generated tests and programming languages performance with Bayesian statistics. Compared to the baseline frequentist statistics, the Bayesian re-analysis violates fewer statistical assumptions and produces more meaningful insights (McElreath, 2018).

While this example demonstrates the improvement of conclusion validity for a single piece of evidence, the same principle translates to variance theories. However, optimizing the conclusion validity of a variance theory requires to systematically compare methods of statistical inference and selecting the one that fits the properties of the data best. At the current point in time, research synthesis offers no advice on selecting an optimal statistical inference method. Quite the opposite, as some synthesis techniques (like AD meta-analysis) are even locked into a frequentist statistical framework.

2.3.4 Construct validity

The last dimension of validity typically discussed in experimentation research is *construct validity*. Construct validity describes the degree to which measurements and instruments reflect a construct of interest, i.e., a value that cannot be directly measured (Cook and Campbell, 1979). Construct validity is the only dimension of validity where a variance theory does not exceed, but is rather limited by the individual pieces of evidence. In fact, varying construct validity of individual pieces of evidence even questions their synthesizability into one variance theory. For example, consider two studies about the impact of TDD on developer productivity. If the first study operationalizes productivity as “number of lines of code written in a time frame” and the second as “self-reported assessment of productivity,” the dependent variables of the two pieces of evidence differ so vastly that a synthesis of the two effect strengths into one variance theory would not be meaningful from a quantitative perspective.¹

¹ Qualitatively speaking, this form of triangulation can be valuable, but not for the quantification of an effect strength.

3 Goal and method

Variance theories produced via research synthesis have the potential to be of greater internal, external, and conclusion validity than individual pieces of evidence. Yet, as discussed in Section 2.3, several shortcomings to the state-of-the-art in research synthesis limit this potential. In this work, we propose a framework that systematically improves the external, internal, and conclusion validity of a variance theory about a phenomenon. As such, it serves as a tool for research communities to orchestrate effective and constructive collaboration across individual studies.

Given the declared goal and the availability of partial methods (like meta-analysis), *method engineering* (Brinkkemper, 1996)—in particular: method assembly (Brinkkemper et al., 1998)—becomes a viable research strategy. Method engineering in SE was introduced by Brinkkemper as “the discipline to construct new methods from parts of existing methods, called method fragments” (Brinkkemper et al., 1998). The three remediable threats to validity of variance theories elaborated in Section 2.3 can be individually addressed with the following, already existing method fragments (Brinkkemper et al., 1998):

- The threat to **external validity** can be addressed through replication studies (Juristo and Gómez, 2012; Gómez et al., 2014; Nosek and Errington, 2020) and meta-analysis (Yang, 2002; Miller, 2000; dos Santos and Travassos, 2020).
- The threat to **internal validity** can be addressed via statistical causal inference (Siebert, 2023; Pearl, 2010), particularly model comparison (McElreath, 2018; Furia et al., 2022; Cinelli et al., 2024).
- The threat to **conclusion validity** can be addressed via statistical reanalysis (Nosek and Errington, 2020; Frattini et al., 2024).

As discussed in Section 2.3.4, given the absence of a proven method fragment to address the threat to construct validity, we do not include it in our framework. Efforts to remedy threats to construct validity (Ralph and Tempero, 2018) are commendable but remain heavily debated (Sjøberg and Bergersen, 2022). To the best of our knowledge, there is no reliable method for systematically assessing the construct validity of an operationalization, and hence, no method to involve in our process theory. The implications of this are minimal for variance theories. It only implies that these variance theories quantify the relationships between *operationalizations*, not *concepts*. While some scholars claim that “theories are formulated at the conceptual level” (Sjøberg and Bergersen, 2022), we neither found evidence for this claim nor see any reason for it. While theories on a conceptual level certainly expand their scope, this elevation comes at the cost of construct validity. And as long as there is no reliable method for assessing construct validity, we decide this tradeoff in favor of validity over scope. Consequently, we consider a valid conclusion about an effect between operationalizations no less valuable than between concepts. Still, we discuss the full implication of this limitation and potential ways to overcome it in Section 6.3.

The method assembled from the three eligible method fragments is on a conceptual level of abstraction and on a method

level of granularity (Brinkkemper et al., 1998). We specify the method from a process perspective that models the relationships and order of involved stages and activities. This serves as a guide to a researcher conducting synthesis how to conduct the process.

4 Framework for research synthesis

Our framework for managing variance theories consists of a definition of evidence in Section 4.1 and, based on it, a framework of relationships between two pieces of evidence in Section 4.2. Section 4.3 then describes how these allow describing the synthesis and evolution of a variance theory.

4.1 Evidence

To describe the relationship between evidence, we first need to formalize the concept of evidence. We define a piece of empirical, quantitative evidence e as a tuple $e := E(h, d, m)$ consisting of three components.

- **Hypothesis h :** A hypothesis consisting of variables and (assumed) causal relationships between those variables. For example, $h_1 : x \rightarrow y$ defines the hypothesis h_1 as variable x causally influencing variable y .
- **Data d :** A record of observations of all variables contained in h . An eligible dataset d_1 for h_1 requires observation for both variables x and y .
- **Method m :** An analysis method that processes the data d under the hypothesis h to produce a conclusion.

Hypotheses are networks of variables and relationships among them. As such, they can be visualized via directed, acyclic graphs (DAGs) (Elwert, 2013) as shown in Figure 1. In these DAGs, nodes represent variables, and directed edges represent assumed causal relationships. Figure 1a shows a graphical representation of a simple, two-variable hypothesis h_1 . More often, though, manuscripts present such simple hypotheses textually. This often takes the form of a verbose null hypothesis, e.g., “ x has no effect on y ” or “There is no significant difference in values of y for different values of x .” More complex hypotheses involving more variables and relationships like Figures 1c, d require graphical representation but are rare in SE research (Siebert, 2023).

The two colored nodes in the four DAGs represent the main *phenomenon* of interest, i.e., the independent variable or *exposure* (colored red) and the *outcome* or response variable (colored cyan). The goal of a piece of empirical, quantitative evidence is to estimate the average causal effect (ACE) of the main exposure(s) on the dependent outcome variable. Additional variables (colored gray) may be relevant to the hypothesis but not part of the main phenomenon under study. For example, Figure 1b implies that z also affects y , i.e., $z \rightarrow y$, in addition to x . Therefore, the same phenomenon of interest can be described in multiple different hypotheses.

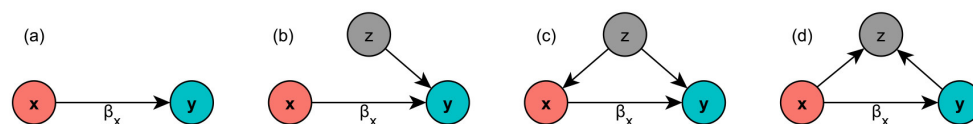


FIGURE 1
DAGs representing (a) a hypothesis between variables x and y , and three revisions introducing a third variable z as (b) a covariate, (c) a confounder, and (d) a collider.

All analysis methods m require deriving a *statistical model* from the causal model—i.e., the hypothesis h (McElreath, 2018). A statistical model typically consists of a regression model, i.e., a specified, often linear relationship between one or more predictors and the outcome variable. In the case of simple, two-variable hypotheses, this boils down to regressing the outcome on the only predictor. For example, the statistical model derived from h_1 in Figure 1a would be $y \sim x$. In the case of more complex hypotheses, one must select a subset of independent variables to condition on, called the *adjustment set*. This subset should maximize the precision of the estimation of the effect of x on y and, on the other hand, ensure that the causal effect is not confounded (Cinelli et al., 2024). For example, the statistical model derived from Figure 1c would be $y \sim x + z$ where z is included to de-confound the effect of x on y . The statistical model of Figure 1d would be $y \sim x$ since including z —called a *collider*—would confound x on y (McElreath, 2018).

The appropriateness of analysis methods depends on the complexity of the hypothesis and the properties of the variables. For simple, two-variable hypotheses consisting of one independent and one dependent variable, most scholars resort to null hypothesis significance tests (NHSTs) like the Student's t -test or its variants (Wagner et al., 2020). Here, the choice depends on whether the dependent variable seems to follow a normal distribution and whether the data is paired or not. For more complex hypotheses involving more than one independent variable, scholars tend to apply linear regression models with multiple predictors (Vegas et al., 2015).

Applying the analysis method m to the data set d based on the causal model implied by the hypothesis h produces the piece of evidence e that offers a conclusion. The nature of the conclusion depends on the analysis method. For example, NHSTs propose a p-value that scholars commonly compare with an arbitrary significance level α to determine whether the independent variable evokes a statistically significant difference in the dependent variable. For linear regression models, the conclusion takes the form of coefficients (e.g., β_x in Figure 1a, representing the strength of the impact of x on y). From these coefficients, one can additionally calculate confidence intervals for each independent variable. If the confidence interval of a variable is not consistent with 0, i.e., it does not intersect 0, then the variable can be considered to have a significant impact on the dependent variable.

4.2 Relationship between evidence

Variance theories emerge from the synthesis of multiple pieces of empirical, quantitative evidence, which increases their

validity and abstracts from passing trends (Hannay et al., 2007). To accommodate research synthesis that goes beyond the meta-analysis of a homogeneous set of primary studies (Pickard et al., 1998), the relationship between two pieces of evidence needs to be clear. Figure 2 visualizes process of synthesizing two or more pieces of evidence, which allows showing the evolution of evidence over time. Starting from an initial piece of evidence $e_1 = E(h_1, d_1, m_1)$, we consider three types visualized as paths in Figure 2. We explain each type by referring to the three branches and the steps within them (to which we refer by their code, e.g., R1 for “conducting a replication”).

4.2.1 Replication

The most commonly known relationship between two pieces of empirical evidence in SE research is via *replication* (left branch colored yellow in Figure 2). A replication subscribes to the same causal hypothesis h_1 and uses the same analysis method m_1 but collects a different data set d_2 to produce a new piece of evidence $e_2 := E(h_1, d_2, m_1)$ (R1). The conclusion derived from the replication e_2 is compared with the conclusion of the original piece of evidence e_1 to check for agreement (R2). Checking for agreement depends on the nature of the conclusion that the analysis method m_1 produces. If m_1 is a type of hypothesis test that produces a p-value, this check is referred to as an *aggregation of p-values* (Santos et al., 2018) via Fisher's or Stouffer's method (Borenstein et al., 2021). If m_1 is a type of regression model that produces confidence intervals of coefficients, then the check boils down to assessing whether the confidence intervals overlap in a Forest plot (Lewis and Clarke, 2001), or AD or IPD-S meta-analysis techniques (Santos et al., 2018). Still, determining the agreement between replications is non-trivial (Destefanis et al., 2026).

If the conclusions agree, the two pieces of evidence e_1 and e_2 can be synthesized to a variance theory $V = E(h_1, \{d_1, d_2\}, m_1)$. This variance theory has an improved external validity over e_1 as the replication shows that the conclusion also in both contexts d_1 and d_2 . However, in case the conclusions disagree, SE literature offers little advice on how to reconcile these results. The disagreeing conclusions indicate a *causal leak* (Martínez-Sánchez et al., 2024), i.e., that at least one variable that would explain the difference between e_1 and e_2 is missing from h_1 . A causal leak requires a *revision* of the original hypothesis.

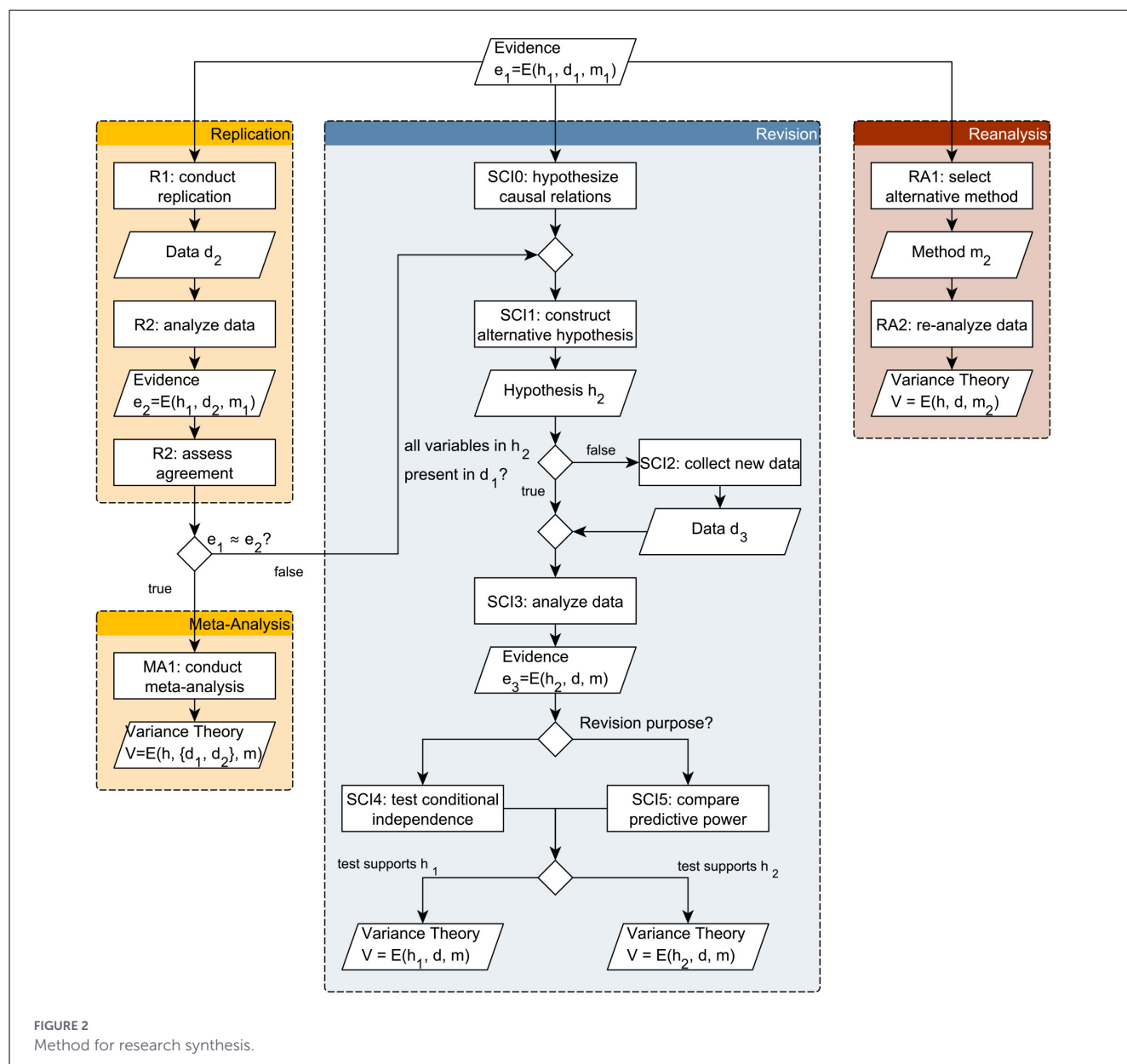


FIGURE 2
Method for research synthesis.

4.2.2 Revision

A new piece of evidence acts as a *revision* of another when the hypothesis h_1 is revised to a competing hypothesis h_2 (middle branch colored blue in Figure 2). This hypothesis supposedly explains the phenomenon under study—which produced the data d_1 and d_2 —better and, therefore, has greater internal validity. The competing network can include new or discard existing variables or can—alternatively or additionally—propose different causal relationships between variables. Only the variables belonging to the phenomenon under study (i.e., the main exposure and outcome) need to remain included. Otherwise, the new hypothesis pertains to a different phenomenon. Figures 1b–d visualize revisions of Figure 1a as they contain the variables of the main phenomenon under study (x and y) but include an additional variable (z) and different relationships.

Disagreeing conclusions from a replication but also emerging qualitative evidence (SCI0) may trigger a revision. For example, a qualitative study might suggest that the variable z also influences y even before observing disagreeing conclusions from replications. Proposing a new hypothesis h_2 may also require collecting a new data set d_3 if h_2 contains variables not recorded in d_1 (SCI2). In the abstract example, a new data set d_3 that records both z in addition to x and y is necessary.

With the new piece of evidence e_3 produced under the proposed hypothesis h_2 (SCI3), the competing hypotheses are evaluated by model comparison. The type of comparison depends on the purpose of the revision. Revisions can serve two different purposes (Cinelli et al., 2024; Navarro, 2019): increasing the precision of the effect estimation, and deconfounding the effect estimation. The first purpose is to increase the precision of estimating the effect β_x of x on y . For example, involving

an additional, independent variable z with an assumed causal relationship $z \rightarrow y$ (see Figure 1b visualizing such an alternative hypothesis h_2) may increase the precision of the estimate of the ACE (Cinelli et al., 2024). By including z in the statistical model, some of the variance previously attributed to β_x will be explained by the influence of z , increasing the precision of the estimation of β_x . The second possible purpose of a revision is to deconfound the estimation of the effect β_x of x on y (McElreath, 2018). This is particularly relevant to phenomena studied in observational, not experimental, settings where the exposure is not a true “independent” variable but can be influenced by factors other than random assignment. A confounder could be a common cause as visualized in Figure 1c where variable z impacts both x and y , therefore biasing the direct effect of x on y . The adjustment set of this hypothesis includes z as a predictor of y to de-confound the effect of x on y (McElreath, 2018).

To evaluate a revision aiming at de-confounding, testable implications in the form of *independencies* and *conditional independencies* are derived from the hypotheses (McElreath, 2018) (SCI4). For example, according to h_3 in Figure 1c, both x and y depend on z . Conversely, h_1 in Figure 1a does not include this claim and implies that x and y are independent of z . Additionally, h_3 implies that the strength of the ACE of x on y changes when conditioning on z via deconfounding, which h_1 does not imply. Correlational analyses on the available data set d_3 can confirm or refute these assumed independencies (McElreath, 2018). Comparing the statistical model $y \sim x$ (derived from h_1) with $y \sim x + z$ (derived from h_3) produces two estimates of the ACE of x on y . If the ACE is the same, then the effect of x on y is independent of z and h_1 represents the causal relations of the phenomenon under study better. If the ACE is different, then the effect depends on z , and h_3 is more valid. Consequently, the internal validity of h_3 exceeds the one of h_1 and can be considered the currently superior causal model to explain the phenomenon under investigation.

To evaluate a revision aiming at increasing the precision, the out-of-sample predictive power of the two models is compared (SCI5) via an appropriate information criterion (McElreath, 2018). Metrics like the Akaike Information Criterion (AIC) (Akaike, 2011) or leave-one-out cross-validation (LOO) can be applied depending on the analysis method (Vehtari et al., 2017; Magnusson et al., 2020). These metrics assign scores to competing hypotheses and infer which predicts the observed data best. The model with the greatest predictive power is assumed to be more internally valid.

While both purposes of revisions aim to strengthen the internal validity of a hypothesis, the respective evaluation that decides the comparison is not interchangeable. Model comparison used to determine the hypothesis with the greater out-of-sample predictive power is not fit when aiming to deconfound evidence, as confounded evidence may very well exhibit a greater predictive power than a deconfounded one (McElreath, 2018). Consequently, the distinction of purpose when conducting a revision is imperative for the choice of evaluation method.

At every point in time, the hypothesis that shows the greatest internal validity should be the one that future studies should subscribe to. This means that all future studies investigating the phenomenon of x and y should (1) record all variables involved in the hypothesis that are part of its adjustment set, and (2) analyze

the collected data under the assumption of the internally most valid hypothesis. The resulting variance theory optimizes the internal validity of the claim about the phenomenon.

Some scholars in SE research would consider revisions a type of replication, but only provide very approximate guidelines to reconcile the difference in hypotheses during synthesis (Gómez et al., 2014). Changing some elements of the experimental design in a revision is sometimes called a reproduction, differentiating it from a similar replication (Cartwright, 1991; Juristo and Gómez, 2012). To clearly distinguish the purpose of revisions from replications and emphasize that their decision process is dissimilar, we use the term “revision” instead of “replication.”

4.2.3 Reanalysis

The least commonly discussed relationship between pieces empirical evidence in SE research is the *reanalysis* of existing data (right branch colored red in Figure 2). Reanalysis—sometimes also referred to as a “test for robustness” (Nosek and Errington, 2020)—describes the application of a different analysis method m_2 to the same data d under the same causal assumptions h (Gómez et al., 2010). In special cases, however, reanalysis may also be necessitated by a revision. For example, extending a hypothesis to include two instead of one independent variable will make analysis methods that only operate with one independent variable (e.g., a t -test) ineligible and necessitate more complex ones (e.g., a linear model).

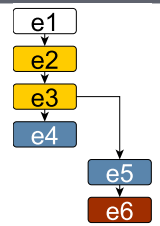
Reanalyses are mostly driven by adapting more advanced methods from other disciplines (e.g., statistics, or evidence-based medical research) (RA1). One instance of this type of evolution is the ongoing endeavor to abandon simple NHSTs for more advanced Bayesian data analysis (Furia et al., 2019). Reanalyses increase the conclusion validity of the evidence by revising statistical assumptions (Wohlin et al., 2012). The decision of which analysis method to prefer over another is often based on intricate statistical comparisons (McElreath, 2018), which SE researchers usually adapt and do not conduct themselves. Re-analyzing existing data sets under given causal hypotheses (RA2) produces a variance theory with greater conclusion validity.

4.3 Evolution

The definition of a piece of evidence in Section 4.1 and how two pieces can relate to each other in Section 4.2 allow modeling the incremental evolution of research about a phenomenon. This yields an *evidence evolution graph* (EEG) which we illustrate in Table 1 and describe as an example in this section. As a starting point, an original study provides a piece of evidence e_1 (written $e1$ in the EEG for better readability) by applying the analysis method m_1 (e.g., a linear model) to data set d_1 under the causal assumption h_1 (defined as $x \rightarrow y$).

All subsequent pieces of evidence e_n investigate the same phenomenon $x \rightarrow y$ but with specific relationships to the original studies. e_2 and e_3 are replications of e_1 , i.e., they subscribe to the same hypothesis h_1 and use the same analysis method m_1

TABLE 1 Exemplary evidence evolution graph.

EEG	Evolution type	Hypothesis	Data	Analysis	Conclusion
	(Original study)	$h_1 : x \rightarrow y$	d_1	m_1	$CI_x = [0.2, 0.6]$
	Replication	h_1	d_2	m_1	$CI_x = [0.15, 0.7]$
	Replication	h_1	d_3	m_1	$CI_x = [-0.4, 0.2]$
	Revision	$h_2 : x \rightarrow y; x \leftarrow z \rightarrow y$	d_4	m_1	z has no effect
	Revision	$h_3 : x \rightarrow y; x \leftarrow a \rightarrow y$	d_5	m_1	a has an effect
	Reanalysis	h_3	d_5	m_2	$CI_x = [-0.2, 0.2]$

but collect different data sets d_2 and d_3 . However, while the conclusion drawn from e_2 largely agrees with e_1 on a positive effect ($[0.2, 0.6] \approx [0.15, 0.7]$), e_3 disagrees ($[0.2, 0.6] \not\approx [-0.4, 0.2]$). The disagreement of conclusions necessitates a revision. e_4 revises the previous hypothesis by proposing h_2 , according to which the variable z confounds the effect $x \rightarrow y$. However, if the conclusion shows that z has no effect, i.e., it does not explain the differences that set e_3 apart from e_1 and e_2 , then this revision does not have the intended effect. h_2 has no greater internal validity than h_1 and is, hence, not worth pursuing further, constituting a “dead end” of the evidence evolution. However, another hypothesis h_3 postulating that instead variable a acts as the confounder might explain the differences previously observed, elevating it to the hypothesis with the greatest internal validity. Future replications should subscribe to h_3 when contributing evidence to the phenomenon $x \rightarrow y$ and make sure to always collect data about a as well to control the confounding. Finally, a reanalysis e_6 might keep both h_3 and d_5 but use a different analysis method m_2 (e.g., a generalized linear model instead of a linear model) if the outcome variable y is shown to fit a different distribution better.

5 Application

To demonstrate the use of the proposed framework, we apply it to two active areas of SE research: test coverage (Section 5.1) and requirements quality (Section 5.2). Both areas implicitly aim at producing variance theories, i.e., quantifying how one factor influences another. The application shows how the use of the framework can relate individual pieces of evidence to a holistic perspective, and simultaneously reveal issues of research synthesis in the absence of a framework.

5.1 Test coverage affecting test suite effectiveness

Determining the effectiveness of a test suite, i.e., its fault detection capability, has been a long standing goal of the testing community. Because the actual fault detection capability can only be determined in hindsight, the community proposed test adequacy criteria like *coverage* metrics to predict the effectiveness of a test suite (Chen et al., 2020). Goodenough and Gerhart (1975) postulated that adequacy criteria like coverage are meaningful

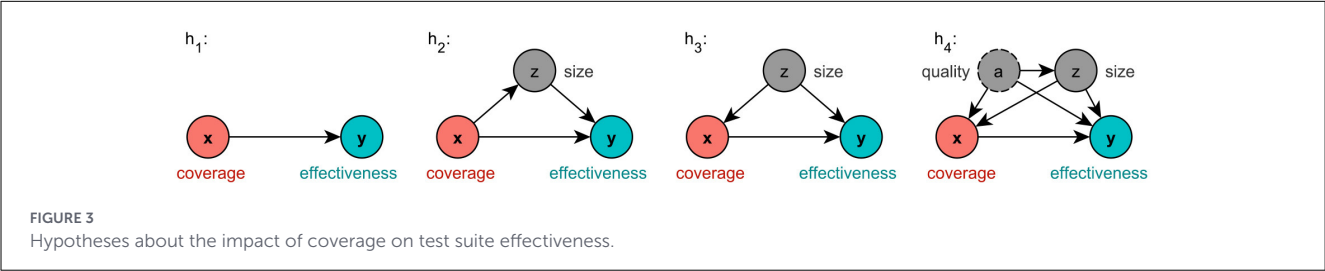
proxies for test reliability. This sparked a long empirical investigation of the claim. Table 2 applies the previously introduced framework to a selection of notable contributions to this line of research.

Frankl and Weiss (1993) tested the hypothesis that a test adequacy criterion like coverage affects test suite effectiveness, denoted as h_1 in Table 2 and visualized in Figure 3. Applying logistic regression to nine programs (e_1), only four of these programs showed a correlation between coverage and effectiveness in fault-detection, rendering the piece of evidence inconclusive. Andrews et al. (2006) replicated the experiment on the industrial-scale `space.c` data set but arrived at the conclusion e_2 that coverage does have a positive effect on effectiveness. To reconcile the contradiction between e_1 and e_2 , the authors additionally conjecture that the *size* of a test suite may mediate the effect: test suites with greater coverage are also likely to be larger in size (i.e., contain more test cases), which could also improve the effectiveness. Based on this revised hypothesis (see h_2 in Figure 3), the authors produced evidence e_3 supporting that both coverage and size have an effect on effectiveness. However, Namin and Andrews (2009) later revise that claim again. In the absence of a framework for causal inference, they claim that it is equally likely that size instead acts as a confounder (see h_3 in Figure 3). However, when comparing h_3 to h_1 , the authors do so by evaluating the model fit via the R^2 score of the two regression models. As elaborated in Section 4.2.2, a revision for the purpose of de-biasing the effect of the phenomenon must be evaluated via testing conditional independencies, not model fit, because the model fit of a biased model could very well exceed the model fit of a de-biased one (McElreath, 2018). In other words: the statistical model derived from h_1 which does *not* control for the confounder size suggested by h_3 may very well have an equal or better model fit. Hence, Namin and Andrews (2009) arrive at an inconclusive piece of evidence where both hypotheses seem equally internally valid. However, Inozemtseva and Holmes (2014) later replicate this revision but with the correct data analysis method: In their evidence e_5 , they compare h_1 with h_3 on 5 Java programs by visualizing the association between *coverage* and *effectiveness*, once controlling *size* (h_3) and once not (h_1). The results clearly show that once controlling *size* the association between *coverage* and *effectiveness* diminishes, supporting h_3 in that *size* acts as a confounder.

Yet, when Gopinath et al. (2014) replicated the study on 250 Java projects, they came to opposite conclusions in e_6 : When controlling *size*, the association between *coverage* and *effectiveness* persisted. To reconcile the once again contradicting replications,

TABLE 2 Evidence evolution about the effect of coverage on mutation scores.

EEG	Type	Ref.	Hyp.	Data	Analysis	Conclusion
e1 e2 e3 e4 e5 e6 e7	(Original)	(Frankl and Weiss, 1993)	h_1	9 programs	Regression	Inconclusive
	Replication	(Andrews et al., 2006)	h_1	space.c	Regression	Positive correlation
	Revision	(Andrews et al., 2006)	h_2	space.c	Regression	Positive correlation
	Revision	(Namin and Andrews, 2009)	h_3	7 programs	Regression	Inconclusive
	Replication	(Inozemtseva and Holmes, 2014)	h_3	5 Java programs	Visualization	Size acts as a confounder
	Replication	(Gopinath et al., 2014)	h_3	250 Java projects	Regression	Size does not confound
	Revision	(Chen et al., 2020)	h_4	-	-	-



the study of [Chen et al. \(2020\)](#)—which included authors from both camps ([Inozemtseva and Holmes, 2014](#); [Gopinath et al., 2014](#))—propose another revision (see h_4 in [Figure 3](#)): While *size* may indeed confound the effect of *coverage* on *effectiveness*, there may be yet another factor at play: the overall *quality* of the test cases, which the authors explain via “developer’s desire to write effective tests” ([Chen et al., 2020](#)). This overall quality may result both in more test cases (*quality* $\rightarrow$ *size*), greater coverage (*quality* $\rightarrow$ *coverage*), and more effective test cases (*quality* $\rightarrow$ *effectiveness*). However, because the factor *quality* was unobserved (indicated by the dashed lines around variable *a* in h_4 in [Figure 3](#)), controlling *size* did not sufficiently de-bias the effect of the phenomenon of interest, which potentially explains the contradictory results in e_5 and e_6 . Sadly, the authors did not provide empirical evidence for h_4 . However, they conclude with the recommendation “that future studies should explicitly state their causal models and assumed underlying processes, which forces a clear statement of scientific questions and enables reasoning about whether the proposed experiments and analyses can answer that question” ([Chen et al., 2020](#)). This need is met by embedding causal modeling in the overall research synthesis framework.

5.2 Passive voice affecting domain model completeness

The application of the proposed framework to the field of requirements quality allows for even more nuanced insights. Since authors of the manuscript at hand were also involved this research agenda ([Frattini et al., 2025, 2024](#)), we can demonstrate how the application of the framework reveals several issues with research synthesis. We show both the evolution of evidence as published (Section 5.2.1) and how it *should* have been conducted (Section 5.2.2).

The research is centered around the question whether the use of *passive voice* when specifying a requirement impacts the *completeness of domain models* resulting from it ([Femmer et al., 2014](#)). For example, when transforming the requirement “All transaction details shall *be obtained* from the Statement Database” into a domain model—i.e., a boxes-and-lines diagram of entities and associations described in the requirement—it is unclear who is responsible for “obtaining transaction details.” The active formulation “*The backend system shall obtain* all transaction details from the Statement Database.” instead makes the association explicit. Variance theories about the impact of passive voice on the completeness of domain models (e.g., “Using passive voice increases makes incomplete domain models 45% more likely”) would help practitioners decide whether it is worth addressing and removing the hypothesized quality defect ([Frattini et al., 2023](#)).

5.2.1 Evolution as published

[Femmer et al. \(2014\)](#) contributed the first piece of empirical evidence e_1 about the impact of passive voice on domain model completeness. The latter construct was measured by counting the number of missing actors, objects, and associations resulting from requirements written in either active or passive voice. In the following, we focus on the effect on *missing associations* (see h_1 in [Figure 4](#)) and how evidence about it evolved (see [Table 3](#)).

In a parallel-design controlled experiment with 15 university students generating domain models from seven requirements each, the application of the Mann–Whitney U test resulted in a statistically significant p-value of 0.001 and an effect size measured via Cliff’s $\delta = 0.75$. The authors concluded that the use of passive voice has a strong effect on the number of missing associations from domain models. [Frattini et al. \(2024\)](#) revisited the claim by (1) revising the hypothesis to h_2 and (2) reanalysing the data with a Bayesian model instead of a frequentist significance test

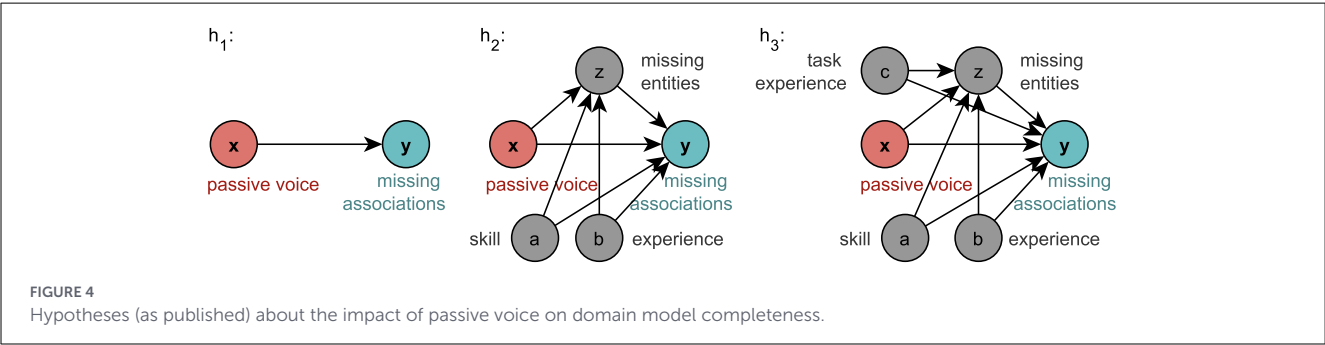


TABLE 3 Evolution of the variance theory on the impact of passive voice on domain modeling.

EEG	Evolution type	Ref.	Hyp.	Data	Analysis method	Conclusion
e1	Original study	(Femmer et al., 2014)	h_1	d_1 : 15 students	m_1 : Mann–Whitney U test	$p = 0.001$
e2	Revision & reanalysis	(Frattini et al., 2024)	h_2	d_1	m_2 : Bayesian model	$[-0.17, \sim 0.49, +0.34]$
e3	Replication	(Frattini et al., 2025)	h_1	d_2 : 25 participants	m_1 : Wilcoxon signed-rank test	$p = 0.025$
e4	Revision & replication	(Frattini et al., 2025)	h_3	d_2	m_2 : Bayesian model	$[-0.25, \sim 0.30, +0.45]$

(cf. second row of Table 3). The latter was motivated by the improved conclusion validity of Bayesian models (Furia et al., 2019; McElreath, 2018). The former was motivated by several conjectures:

1. Domain models derived from passive-voice requirements may miss associations because they are missing entities (i.e., actors or objects) in the first place (i.e., *passive voice* $\rightarrow$ *missing entities* $\rightarrow$ *missing associations*).
2. The self-reported academic and industrial *experience* may influence the number of missing elements.
3. The subjects participating in the study may have different *skill*, operationalized as a random effect in the analysis per participant ID.

The resulting evidence e_2 retains only a very weak effect of passive voice: only in 34% of all cases, passive voice causes more missing associations, while in 49% they remain the same and in 17% of all cases, passive voice sentences cause even fewer missing associations.

In a second follow-up study, Frattini et al. (2025) contributed two pieces of evidence e_3 and e_4 (cf. third and forth row of Table 3). Based on a new data set d_2 collected from a within-subject design controlled experiment with 25 mostly industrial participants, they first replicated the original study. Using the Wilcoxon signed-rank test (as the data in d_2 was paired, other than in d_1) under the assumption of h_1 , evidence e_3 supported e_1 with $p = 0.025$ (assuming a significance threshold of $\alpha = 0.05$). Additionally, they conducted a joint revision and replication, i.e., e_4 further revises the hypothesis by challenging h_2 with h_3 (see Figure 4) and uses d_2 instead of d_1 . The revised hypothesis was motivated h_3 by even more factors that potentially influence the outcome variable, like prior experience in domain modeling (labeled *task experience* in Figure 4). e_4 agrees with e_2 in that the impact of passive voice on the number of missing associations from domain models is not strictly negative, as the likelihood of missing an association remains below 50%. However, e_4 suggests that the negative impact is more likely than assumed by e_2 (45% instead of 34%).

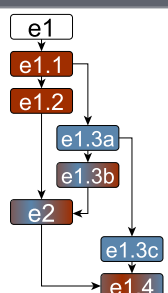
5.2.2 Evolution as supposed to publish

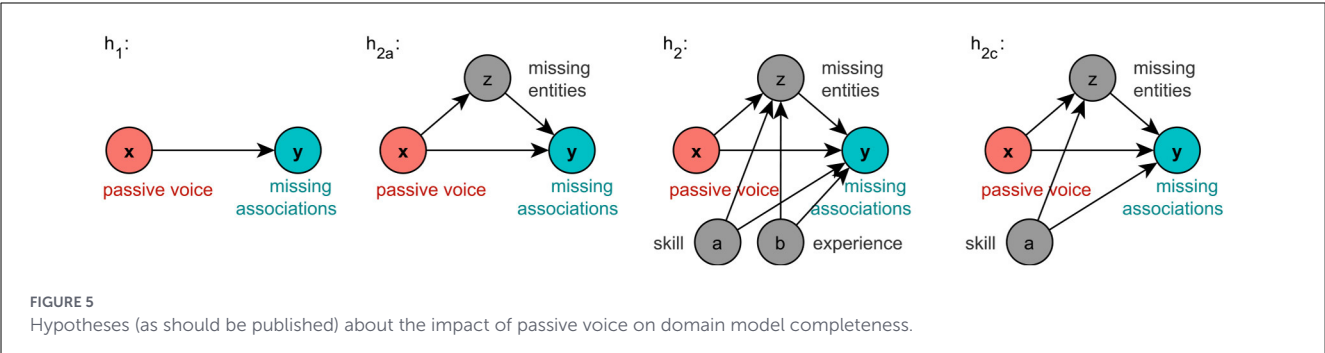
In the absence of an explicit framework to relate individual pieces of empirical evidence, the otherwise coherent research about the impact of passive voice on domain model completeness does not arrive at a proper variance theory. Already the first follow-up study (Frattini et al., 2024) conflates a revision (i.e., changing the hypothesis from h_1 to h_2) with a reanalysis (i.e., changing the analysis method from a Mann–Whitney U test m_1 to a Bayesian model m_2), which makes it impossible to clearly attribute the diverging results to one of the changes. Since we have access to all data and material, we were able to disentangle the implicit steps between e_1 and e_2 , as outlined in Table 4.

Resolving the conflation between the revision and reanalysis of e_2 in relation to e_1 prioritizes the reanalysis, since the revision requires an analysis method accepting more than one predictor (which the Mann–Whitney U test m_1 does not allow). An intermediate reanalysis $e_{1.1}$ where the frequentist test is replaced by a standard linear regression model (Lindeløv, 2019) eases the comparison with the Bayesian reanalysis. The intermediate step shows that the pure reanalysis $e_{1.2}$ (i.e., e_2 without the revision but using the Bayesian model m_2) largely agrees with the frequentist test, although remaining more cautious. This contradicts the conclusion of the authors of the conflated revision and reanalysis, who attributed the observed difference to the use of Bayesian statistics (Frattini et al., 2024).

Replacing the simple significance test of e_1 with a linear model in $e_{1.1}$ allows systematically revising the hypothesis h_1 to isolate the impact of different additions. The revision of h_1 to h_2 implies two revisions with two different purposes: the addition of the mediator *missing entities* has causal implications (McElreath, 2018) while the addition of covariates like *skill* serves the precision of the estimate (Cinelli et al., 2024). Separating these two isolates their effect. Adding only the mediator *missing entities* (h_{2a} in Figure 5) reduces the confidence interval from $[0.23, 0.84]$ to $[0.17, 0.77]$, but the conclusion remains the same. Assessing the internal validity for

TABLE 4 Decomposed evolution from e_1 to e_2 .

EEG	Evolution type	Ref.	Hyp.	Analysis method	Conclusion
	Original study	(Femmer et al., 2014)	h_1	m_1 : Mann–Whitney U test	$p = 0.001$
	Reanalysis		h_1	$m_{1,1}$: linear model	$ci = [0.23, 0.84]$
	Reanalysis		h_1	m_2 : Bayesian model	$[-0.15, \sim 0.34, +0.51]$
	Revision		h_{2a}	$m_{1,1}$: linear model	$ci = [0.17, 0.77]$
	Revision		h_2	$m_{1,2}$: linear mixed model	$ci = [-0.17, 0.82]$
	Revision/reanalysis	(Frattini et al., 2024)	h_2	m_2 : Bayesian model	$[-0.17, \sim 0.49, +0.34]$
	Revision		h_{2c}	$m_{1,2}$: linear mixed model	$ci = [0.03, 0.92]$
	Reanalysis/revision		h_{2c}	m_2 : Bayesian model	$[-0.14, \sim 0.47, +0.39]$



this revision requires testing conditional independencies. Including *missing entities* as a mediator implies that *missing associations* is not independent of it, which is shown by the authors: missing more entities causes missing more associations (Frattini et al., 2024). As such, h_{2a} has greater internal validity than h_1 .

However, once adding the covariate *skill* operationalized as a subject specific random effect (which necessitates switching the linear model $m_{1,1}$ out for a linear *mixed* model $m_{1,2}$) the confidence interval widens and overlaps with 0, such that the effect can no longer be considered statistically significant. To determine whether the revision from h_{2a} to h_2 has greater internal validity and whether the reanalysis from the linear model to the linear mixed model has greater conclusion validity, two checks are necessary. Deciding greater conclusion validity requires assessing statistical properties of the two pieces of evidence. For instance, the residuals of a linear model should be independent and identically distributed (iid) (West et al., 2022). Both the histograms of the residuals, the QQ-plot, and the Durbin–Watson test [autocorrelation greater than 0 ($p = 0.07$)] suggest that the residuals of $e_{1,1}$ are not iid. Consequently, applying a linear model $m_{1,1}$ may lead to invalid conclusions, and the conclusion validity of a linear mixed model $m_{1,2}$ is greater (West et al., 2022).

Deciding greater internal validity requires assessing predictive power in this case. We calculate the AIC, which applies to both $m_{1,1}$ and $m_{1,2}$ (Vaida and Blanchard, 2005), achieving scores of $AIC(e_{1,3a}) = 249.1$ and $AIC(e_{1,3b}) = 251.2$. The score differential of about two points is considered negligible when interpreting the AIC values (Greenwood, 2021). Hence, there is no strong evidence that h_2 is more internally valid than h_{2a} .

Finally arriving at the evidence e_2 (Frattini et al., 2024) is now consistent from both angles: e_2 acts as a revision of $e_{1,2}$ by

changing only the hypothesis h_1 to h_2 , for which the Bayesian model concludes that the effect of passive voice is smaller (i.e., causing more missed associations in only 34% instead of 51% of all cases). Simultaneously, e_2 acts as a reanalysis of $e_{1,3b}$ by changing the linear mixed model to a Bayesian model. e_2 concludes that in the majority of cases (49%), the use of passive voice does not in- or decrease the number of missed associations from a domain model, which the confidence interval $[-0.17, 0.82]$ overestimates.

However, the decomposed evolution of evidence showed that there is little support for h_2 to be more internally valid than h_1 . Yet, the diagnostics of $e_{1,3b}$ reveal that the inclusion of *skill* as a random effect had a strong impact: While $e_{1,3b}$ only has a marginal R^2 value of 0.184, it has a conditional R^2 value of 0.561, which indicates that the random effects explain a lot more of the variance of the outcome variable than the fixed effects (Nakagawa and Schielzeth, 2013). This is in line with observations in empirical research, that—when analyzing experiments with small sample sizes—the within-subject variance of subjects (e.g., their skill) has much more impact on the outcome than other measured variables (Gelman and Hill, 2007; Baayen et al., 2008). The benefit of the random effects in h_2 may be offset by the number of predictors, as the AIC metric penalizes an increased number of predictors to avoid overfitting (Greenwood, 2021). This suggests the competing hypothesis h_{2c} which retains the provenly effective random effects for participants' *skill* but discards the fixed effects of academic and industrial *experience*. The operationalization of the fixed effect is questionable, as it was simply measured on an ordinal scale with four levels. Therefore, the inclusion of these fixed effects might not benefit the estimation and rather overfit the estimation. The predictive power $AIC(e_{1,3c}) = 241.4$ is significantly—i.e., more than 2 units (Greenwood, 2021)—lower than $AIC(e_{1,3a}) = 249.1$ and $AIC(e_{1,3b}) = 251.2$, suggesting

that it has the greatest internal validity and should have been used subsequently instead of h_2 . Evidence $e_{1.3c}$ concludes that passive voice has an impact of $ci_{e_{1.3c}}(passive) = [0.03, 0.92]$, which is still not consistent with 0 but broader than $ci_{e_{1.3b}}(passive) = [0.17, 0.77]$. Using the same line of reasoning as before, the final piece of evidence $e_{1.4}$ could have combined the internally most valid hypothesis with analysis method of the highest conclusion validity, the Bayesian model m_2 . $e_{1.4}$ concludes that using passive voice in requirements causes fewer missed associations in 14%, equal in 47%, and more in 39% of all cases.

6 Discussion

As shown in the previous demonstrations, the framework proposed in Section 4 assembles method fragments to a coherent and consistent approach to improving the internal, external, and conclusion validity of variance theories. Section 6.1 discusses the general implications of the framework, and Section 6.2 elaborates three concrete use cases by which the framework can improve SE research. However, we also acknowledge limitations in Section 6.3 which necessitate the future work described in Section 6.4.

6.1 Implications

The proposed framework has the ability to impact both the knowledge creation, synthesis, and application phase of empirical research, effectively supporting the complete epistemological life cycle. Most prominently, the framework provides the terminology and process to synthesize individual pieces of empirical evidence about a phenomenon. This has the potential to improve the aggregation and integration of evidence in SE research and clearly identify how the frontier of the body of knowledge evolves. At the same time, the framework also clearly values the contribution of replications, revisions, and reanalyses. This increases the incentive to pursue follow-up research and not abandon empirical investigations after just one single study.

By extension, the synthesis process centered around evidence evolution graphs can also improve the knowledge creation phase. Future studies contributing new evidence to an existing phenomenon will be able to better gauge the current frontier of a body of knowledge, and therefore, inform their design. Proven hypotheses with maximum internal and analysis methods with maximum conclusion validity suggest which variables to collect and how to analyze data recording them.

This produces more coherent research that—after systematic synthesis—also becomes more digestible for practitioners in the form of focused, integrated results with maximum internal, external, and conclusion validity. Such results have a higher chance of application in practice as they are more convincing than isolated one-off studies. As a consequence, empirical SE research has the chance of contributing more practically relevant support for the intended target population of software engineers.

6.2 Use cases

Such implications can be realized via several use cases that improve the state of empirical SE research. We present these use cases in ascending order of complexity, starting with the most accessible one.

6.2.1 Framing new evidence

The framework helps researchers to position a piece of evidence in relation to an existing body of knowledge. It provides a terminology to frame how new studies advance the body of knowledge with respect to existing ones: new evidence can be classified as either a replication, revision, or reanalysis, depending on what parts of the empirical evidence changed in respect to its origin, and the framework encourages changing only one part at a time. More importantly, the framework supports deciding whether this new evidence strengthens or challenges the body of knowledge:

1. **Replications** (i.e., changing the data) where the conclusion
 - (a) agrees with the predecessor *improve the external validity*.
 - (b) disagrees with the predecessor *necessitate a revision*.
2. **Revisions** (i.e., changing the hypothesis) *improve the internal validity* if their purpose is
 - (a) increasing the precision of the ACE estimation and the revision has a higher predictive power.
 - (b) deconfounding the ACE estimation and the revision is more consistent with (conditional) independencies.
3. **Reanalyses** (i.e., changing the analysis method) *improve conclusion validity* if the new analysis method violates fewer statistical assumptions.

6.2.2 Systematic literature reviews

In extension to the prior use case, the framework allows to model the complete evolution of empirical evidence about a phenomenon, as shown in Section 5. This aligns with the original purpose of a systematic literature review (SLR) as originally intended by Kitchenham and Charters (2007). By framing all empirical, quantitative studies investigating one phenomenon using the framework, the evolution of a variance theory about that phenomenon can be described, analyzed, and allows to make explicit and informed decision about future research directions. Applying our framework to an SLR would produce a central EEG as shown in Section 5, which clearly shows which causal assumptions about the phenomenon are best supported (i.e., have the highest internal validity) and which are “dead ends.” This reveals the current empirical knowledge quantitative studies on a phenomenon and can inform future study design.

6.2.3 Informing a follow-up study

Once an EEG of a phenomenon exists, researchers planning to conduct a follow-up study about the phenomenon can consult

the ECG to improve the design of their study. Two aspects revealed by an EEG are particularly relevant. Firstly, the hypothesis with the highest internal validity informs which variables the researchers should collect. For example, in the context of test coverage research (cf. Section 5.1), future research should at least collect data on *test suite size* following Inozemtseva and Holmes (2014). Even better, additionally attempting to measure or control the underlying *quality* aspiration with which the test suites were created could bring empirical evidence to the claim of Chen et al. (2020). Secondly, the analysis method with the highest conclusion validity informs how the researchers should analyze their data. For example, in the context of requirements quality research (cf. Section 5.2), Bayesian data analysis methods have shown to violate fewer statistical assumptions (Frattini et al., 2024), and hence, should be preferred when conducting a new study.

6.3 Limitations

Despite the potential advantages for empirical SE research, the proposal remains subject to several threats and limitations.

6.3.1 Reliance on fragments' validity

Most prominently, the validity of the framework relies on the validity of the method fragments it is assembled from. Since prior research extensively supports the validity of replications (Nosek and Errington, 2020; Gómez et al., 2010), meta-analysis (Yang, 2002; Kitchenham et al., 2020), statistical causal inference (Pearl, 2010; McElreath, 2018), and reanalysis (McElreath, 2018), we deem this threat constrained.

6.3.2 Limitation to quantitative evidence

By design of managing *variance* theories, the proposed framework focuses on integrating quantitative evidence. Qualitative evidence can trigger the revision of a hypothesis (cf. Section 4.2.2), but does not qualify as an individual piece of evidence involved in the synthesis due to its lack of quantification.

6.3.3 Construct validity

The proposed framework circumvents construct validity entirely by focusing on variance theories about effects between *operationalizations*, not concepts. While this eliminates all threats to construct validity, it also limits the applicability of the resulting variance theories. For example, the variance theory regarding test coverage in Section 5.1 pertains to the effect of the operational *test coverage*, not the broader concept of *test adequacy criteria*. Given the lack of a reliable method fragment to improve construct validity (Sjøberg and Bergersen, 2022), the framework contributes no systematic improvement along this dimension of validity.

6.3.4 Quality appraisal

As with any form of secondary study, the quality appraisal of primary studies—i.e. the decision, which primary studies to include in the synthesis and which to exclude—is non-trivial (Kitchenham and Charters, 2007). The proposed framework does not address this issue and only focuses on the relation and comparison of individual pieces of evidence once included. The careful appraisal and selection of primary studies to integrate remains an ongoing challenge.

6.4 Future work and roadmap

The retrospective application of the proposed framework has the potential to provide additional structure to SLRs. Centering SLRs around evidence evolution graphs with explicit hypotheses, data, and analysis methods clearly describes the progress and frontier of research about a phenomenon. For existing SLRs where this information is available, one avenue of future work is their replication using the proposed synthesis framework, similar to Sections 5.1 and 5.2. The main impact of the framework, however, is expected on new SLRs and new empirical evidence contributing to SLRs. We envision a future of empirical SE research with the prospective adoption of the framework to not only synthesize quantitative evidence to variance theories, but also inform research design and dissemination. Future work can further improve this impact.

From the perspective of researchers, the next step is facilitating the adoption of the framework. We encourage researchers to structure their research efforts around a visual EEG that summarizes the evolution of empirical evidence hitherto. Such EEG would act as a central guidance system coordinating future research about the same phenomenon. A natural extension would be an online platform hosting the EEG where new evidence about the same phenomenon could be submitted, integrated, and maintained, similar to how Github acts as a platform enabling the distributed version control of source code repositories. Even further, such platform could enable the automatic synthesis of all pieces of evidence following the same hypothesis and using the same analysis method. This would mark a paradigm shift away from a static, incidental, and manual toward dynamic and continuous synthesis of primary studies.

From the perspective of practitioners, the next step is accessing the synthesized knowledge via these platforms. If the variance theory about a phenomenon synthesized from all available and recent pieces of evidence is available to SE practitioners, SE research could improve the aspired knowledge translation (dos Santos and Travassos, 2020) by presenting incrementally synthesized variance theories rather than unconvincing one-off studies to practitioners (Miller, 2005).

7 Conclusion

Research synthesis can improve knowledge translation by producing variance theories with greater validity than individual pieces of evidence. To extend synthesis practices in SE research

beyond meta-analyses which only apply under specific conditions, we propose a definition of empirical, quantitative evidence and a framework for its evolution. The framework allows to clearly relate two pieces of evidence about a phenomenon and supports deciding which has the greater internal, external, and conclusion validity. Putting multiple pieces of evidence into relation with each other reveals the evolution of empirical research about a phenomenon and indicates its current frontier. We believe that this vision for orchestrating empirical research about a phenomenon of interest will make research progress explicit, support the knowledge generation, synthesis, and application process, and enable productive, distributed research endeavors in our field of SE research. We hope it will help us move beyond isolated islands of individual studies, and toward more coherent, constructive, and reliable research outcomes.

Data availability statement

A preprint of this article was submitted to arXiv, (Frattini, 2026), doi: 10.48550/arXiv.2609.05983 in the computer science/software engineering collection.

Author contributions

JFr: Writing – original draft, Investigation, Writing – review & editing, Methodology, Formal analysis, Project administration, Data curation, Validation, Visualization, Conceptualization. JFi: Conceptualization, Data curation, Writing – review & editing, Validation, Writing – original draft. DF: Validation, Conceptualization, Supervision, Writing – review & editing, Investigation, Writing – original draft. MU: Methodology, Conceptualization, Supervision, Writing – review & editing, Writing – original draft. DM: Writing – review & editing, Conceptualization, Supervision, Methodology, Writing – original draft. RF: Writing – review & editing, Writing – original draft, Conceptualization, Supervision, Validation. RT: Conceptualization, Writing – original draft, Writing – review & editing, Validation, Supervision.

Funding

The author(s) declared that financial support was received for this work and/or its publication. This work was supported by the

KKS foundation through the S.E.R.T. Research Profile project, the GIST project at Blekinge Institute of Technology, and the Chalmers Foundation through the Chalmers Academic Excellence Program.

Acknowledgments

We express our deep gratitude to the participants of the ISERN workshop who critically reviewed this work and helped revise it. Particularly, we thank Stefan Wagner, Sebastian Baltes, Maria Teres Baldassarre, Guilherme Horta Travassos, Jefferson Seide Moller, Andreas Jedlitschka, Stefan Biffl, and Laurie Williams for their constructive feedback.

Conflict of interest

JFi was employed by Netlight Consulting GmbH.

The remaining author(s) declared that this work was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declared that Generative AI was not used in the creation of this manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

- Akaike, H. (2011). "Akaike's information criterion," in *International Encyclopedia of Statistical Science*, ed. M. Lovric (Berlin: Springer), 25. doi: 10.1007/978-3-642-04898-2_110
- Andrews, J. H., Briand, L. C., Labiche, Y., and Namin, A. S. (2006). Using mutation analysis for assessing and comparing testing coverage criteria. *IEEE Trans. Softw. Eng.* 32, 608–624. doi: 10.1109/TSE.2006.83
- Baayen, R. H., Davidson, D. J., and Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *J. Mem. Lang.* 59, 390–412. doi: 10.1016/j.jml.2007.12.005
- Basili, V. R., Shull, F., and Lanubile, F. (1999). Building knowledge through families of experiments. *IEEE Trans. Softw. Eng.* 25, 456–473. doi: 10.1109/32.799939

- Beecham, S., O'Leary, P., Baker, S., Richardson, I., and Noll, J. (2014). Making software engineering research relevant. *Computer* 47, 80–83. doi: 10.1109/MC.2014.92
- Borenstein, M., Hedges, L. V., Higgins, J. P., and Rothstein, H. R. (2021). *Introduction to Meta-Analysis*. Hoboken, NJ: John Wiley Sons. doi: 10.1002/9781119558378
- Brinkkemper, S. (1996). Method engineering: engineering of information systems development methods and tools. *Inf. Softw. Technol.* 38, 275–280. doi: 10.1016/0950-5849(95)01059-9
- Brinkkemper, S., Saeki, M., and Harmsen, F. (1998). "Assembly techniques for method engineering," in *International conference on advanced information systems engineering* (Pisa: Springer), 381–400. doi: 10.1007/BFb0054236
- Budgen, D., Kitchenham, B., and Brereton, P. (2013). "The case for knowledge translation," in *2013 ACM/IEEE international symposium on empirical software engineering and measurement* (Baltimore, MD: IEEE), 263–266. doi: 10.1109/ESEM.2013.41
- Cartaxo, B., Pinto, G., Vieira, E., and Soares, S. (2016). "Evidence briefings: towards a medium to transfer knowledge from systematic reviews to practitioners," in *Proceedings of the 10th ACM/IEEE international symposium on empirical software engineering and measurement* (Ciudad Real), 1–10. doi: 10.1145/2961111.2962603
- Cartwright, N. (1991). Replicability, reproducibility, and robustness: comments on harry collins. *Hist. Polit. Econ.* 23, 143–155. doi: 10.1215/00182702-23-1-143
- Chen, Y. T., Gopinath, R., Tadakamalla, A., Ernst, M. D., Holmes, R., Fraser, G., et al. (2020). "Revisiting the relationship between fault detection, test adequacy criteria, and test set size," in *Proceedings of the 35th IEEE/ACM international conference on automated software engineering* (Melbourne, VIC), 237–249. doi: 10.1145/3324884.3416667
- Cinelli, C., Forney, A., and Pearl, J. (2024). A crash course in good and bad controls. *Sociol. Methods Res.* 53, 1071–1104. doi: 10.1177/00491241221099552
- Ciolkowski, M. (2009). "What do we know about perspective-based reading? An approach for quantitative aggregation in software engineering," in *2009 3rd international symposium on empirical software engineering and measurement* (Lake Buena Vista, FL: IEEE), 133–144. doi: 10.1109/ESEM.2009.5316026
- Ciolkowski, M., and Münch, J. (2005). Accumulation and presentation of empirical evidence: problems and challenges. *ACM SIGSOFT Softw. Eng. Notes* 30, 1–3. doi: 10.1145/1083174.1083178
- Cook, T. D., and Campbell, D. T. (1979). *The Design and Conduct of True Experiments and Quasi-Experiments in Field Settings*. In *Reproduced in part in Research in Organizations: issues and Controversies*. Santa Monica, CA: Goodyear Publishing Company.
- Cook, T. D., Campbell, D. T., and Day, A. (1979). *Quasi-Experimentation: Design and Analysis Issues for Field Settings*. Vol 351. Boston, MA: Houghton Mifflin Boston.
- Cruzes, D. S., and Dyba, T. (2011). Research synthesis in software engineering: a tertiary study. *Inf. Softw. Technol.* 53, 440–455. doi: 10.1016/j.infsof.2011.01.004
- Destefanis, G., Shepperd, M., and Yousefi, L. (2026). "The replication assessment problem in software engineering," in *Proceedings of the 30th international conference on evaluation and assessment in software engineering* (Glasgow).
- Devanbu, P., Zimmermann, T., and Bird, C. (2016). "Belief and evidence in empirical software engineering," in *Proceedings of the 38th international conference on software engineering* (Austin, TX), 108–119. doi: 10.1145/2884781.2884812
- Dieste, O., Fernández, E., Martínez, R. G., and Juristo, N. (2011). "Comparative analysis of meta-analysis methods: when to use which?" in *15th annual conference on evaluation and assessment in software engineering (EASE 2011)* (Durham: IET), 36–45. doi: 10.1049/ic.2011.0005
- Dixon-Woods, M., Agarwal, S., Jones, D., Young, B., and Sutton, A. (2005). Synthesising qualitative and quantitative evidence: a review of possible methods. *J. Health Serv. Res. Policy* 10, 45–53. doi: 10.1177/135581960501000110
- dos Santos, P. S. M., and Travassos, G. H. (2013). On the representation and aggregation of evidence in software engineering: a theory and belief-based perspective. *Electron. Notes Theor. Comput. Sci.* 292, 95–118. doi: 10.1016/j.entcs.2013.02.008
- dos Santos, P. S. M., and Travassos, G. H. (2017). "Structured synthesis method: the evidence factory tool," in *2017 ACM/IEEE international symposium on empirical software engineering and measurement (ESEM)* (Toronto, ON: IEEE), 480–481. doi: 10.1109/ESEM.2017.68
- dos Santos, P. S. M., and Travassos, G. H. (2020). "Research synthesis in software engineering," in *Contemporary Empirical Methods in Software Engineering*, eds. M. Felderer, and G. H. Travassos (Cham: Springer), 443–474. doi: 10.1007/978-3-030-32489-6_16
- Elwert, F. (2013). "Graphical causal models," in *Handbook of Causal Analysis for Social Research*, ed. S. L. Morgan (Dordrecht: Springer), 245–273. doi: 10.1007/978-94-007-6094-3_13
- Femmer, H., Kučera, J., and Vetró, A. (2014). "On the impact of passive voice requirements on domain modelling," in *Proceedings of the 8th ACM/IEEE international symposium on empirical software engineering and measurement* (Torino), 1–4. doi: 10.1145/2652524.2652554
- Fernández, D. M., and Passoth, J.-H. (2019). Empirical software engineering: from discipline to interdiscipline. *J. Syst. Softw.* 148, 170–179. doi: 10.1016/j.jss.2018.11.019
- Franch, X., Mendez, D., Vogelsang, A., Heldal, R., Knauss, E., Oriol, M., et al. (2020). How do practitioners perceive the relevance of requirements engineering research? *IEEE Trans. Softw. Eng.* 48, 1947–1964. doi: 10.1109/TSE.2020.3042747
- Frankl, P. G., and Weiss, S. N. (1993). An experimental comparison of the effectiveness of branch testing and data flow testing. *IEEE Trans. Softw. Eng.* 19, 774–787. doi: 10.1109/32.238581
- Frattini, J. (2026). *Replication Package for "Replications, Revisions, and Reanalyses: managing Empirical Evidence in Software Engineering"*. Geneva: Zenodo. doi: 10.5281/zenodo.14288346
- Frattini, J., Fucci, D., Torkar, R., and Mendez, D. (2024). A second look at the impact of passive voice requirements on domain modeling: bayesian reanalysis of an experiment. *arXiv [Preprint]*. arXiv:2402.10800. doi: 10.1145/3643664.3648211
- Frattini, J., Fucci, D., Torkar, R., Montgomery, L., Unterkalmsteiner, M., Fischbach, J., et al. (2025). Applying bayesian data analysis for causal inference about requirements quality: a controlled experiment. *Empir. Softw. Eng.* 30:29. doi: 10.1007/s10664-024-10582-1
- Frattini, J., Montgomery, L., Fischbach, J., Mendez, D., Fucci, D., and Unterkalmsteiner, M. (2023). Requirements quality research: a harmonized theory, evaluation, and roadmap. *Requir. Eng.* 28, 1–14. doi: 10.1007/s00766-023-00405-y
- Frattini, J., Torkar, R., Feldt, R., and Furia, C. (2026). "Towards improving the external validity of software engineering experiments with transportability methods," in *Proceedings of the 30th international conference on evaluation and assessment in software engineering* (Glasgow).
- Furia, C. A., Feldt, R., and Torkar, R. (2019). Bayesian data analysis in empirical software engineering research. *IEEE Trans. Softw. Eng.* 47, 1786–1810. doi: 10.1109/TSE.2019.2935974
- Furia, C. A., Torkar, R., and Feldt, R. (2022). Applying bayesian analysis guidelines to empirical software engineering data: the case of programming languages and code quality. *ACM Trans. Softw. Eng. Methodol.* 31, 1–38. doi: 10.1145/3490953
- Gelman, A., and Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. Cambridge: Cambridge University Press. doi: 10.32614/CRAN.package.arm
- Gómez, O. S., Juristo, N., and Vegas, S. (2010). "Replications types in experimental disciplines," in *Proceedings of the 2010 ACM-IEEE international symposium on empirical software engineering and measurement* (Bolzano-Bozen), 1–10. doi: 10.1145/1852786.1852790
- Gómez, O. S., Juristo, N., and Vegas, S. (2014). Understanding replication of experiments in software engineering: a classification. *Inf. Softw. Technol.* 56, 1033–1048. doi: 10.1016/j.infsof.2014.04.004
- Goodenough, J. B., and Gerhart, S. L. (1975). "Toward a theory of test data selection," in *Proceedings of the international conference on reliable software* (Los Angeles, CA), 493–510. doi: 10.1145/800027.808473
- Gopinath, R., Jensen, C., and Groce, A. (2014). "Code coverage for suite evaluation by developers," in *Proceedings of the 36th international conference on software engineering* (Hyderabad), 72–82. doi: 10.1145/2568225.2568278
- Greenwood, M. C. (2021). *Intermediate Statistics with R*. Bozeman, MT: Montana State University.
- Hannay, J. E., Sjöberg, D. I., and Dyba, T. (2007). A systematic review of theory use in software engineering experiments. *IEEE Trans. Softw. Eng.* 33, 87–107. doi: 10.1109/TSE.2007.12
- Harris, M. J., and Rosenthal, R. (1985). Mediation of interpersonal expectancy effects: 31 meta-analyses. *Psychol. Bull.* 97:363. doi: 10.1037/0033-2909.97.3.363
- Hayes, W. (1999). "Research synthesis in software engineering: a case for meta-analysis," in *Proceedings sixth international software metrics symposium (Cat. No. PR00403)* (Boca Raton, FL), 143–151. IEEE. doi: 10.1109/METRIC.1999.809735
- Hosseini, S., Turhan, B., and Gunarathna, D. (2017). A systematic literature review and meta-analysis on cross project defect prediction. *IEEE Trans. Softw. Eng.* 45, 111–147. doi: 10.1109/TSE.2017.2770124
- Inozemtseva, L., and Holmes, R. (2014). "Coverage is not strongly correlated with test suite effectiveness," in *Proceedings of the 36th international conference on software engineering* (Hyderabad), 435–445. doi: 10.1145/2568225.2568271
- Juristo, N., and Gómez, O. S. (2012). *Replication of Software Engineering Experiments*. Berlin: Springer, 60–88. doi: 10.1007/978-3-642-25231-0_2
- Juristo, N., and Vegas, S. (2011). The role of non-exact replications in software engineering experiments. *Empir. Softw. Eng.* 16, 295–324. doi: 10.1007/s10664-010-9141-9
- Kitchenham, B. (2008). The role of replications in empirical software engineering—a word of warning. *Empir Softw Eng*, 13:219–221. doi: 10.1007/s10664-008-9061-0
- Kitchenham, B., and Charters, S. (2007). *Guidelines for Performing Systematic Literature Reviews in Software Engineering*. Technical report, Keele University/University of Durham.
- Kitchenham, B., Madeyski, L., and Brereton, P. (2020). Meta-analysis for families of experiments in software engineering: a systematic review and reproducibility and validity assessment. *Empir. Softw. Eng.* 25, 353–401. doi: 10.1007/s10664-019-09747-0

- Kitchenham, B. A., Dyba, T., and Jorgensen, M. (2004). "Evidence-based software engineering," in *Proceedings. 26th international conference on software engineering* (Edinburgh: IEEE), 273–281. doi: 10.1109/ICSE.2004.1317449
- Lewis, S., and Clarke, M. (2001). Forest plots: trying to see the wood and the trees. *BMJ* 322, 1479–1480. doi: 10.1136/bmj.322.7300.1479
- Lindelo, J. (2019). *Common Statistical Tests Are Linear Models (Or: How to Teach Stats)*. Available online at: <https://lindeloev.github.io/tests-as-linear/> (Accessed November 11, 2024).
- Lo, D., Nagappan, N., and Zimmermann, T. (2015). "How practitioners perceive the relevance of software engineering research," in *Proceedings of the 2015 10th joint meeting on foundations of software engineering* (Bergamo), 415–425. doi: 10.1145/2786805.2786809
- Magnusson, M., Vehtari, A., Jonasson, J., and Andersen, M. (2020). "Leave-one-out cross-validation for bayesian model comparison in large data," in *International conference on artificial intelligence and statistics* (Palermo: PMLR), 341–351.
- Martínez-Sánchez, Á., Arranz, G., and Lozano-Durán, A. (2024). Decomposing causality into its synergistic, unique, and redundant components. *Nat. Commun.* 15:9296. doi: 10.1038/s41467-024-53373-4
- McElreath, R. (2018). *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*. Boca Raton, FL: Chapman and Hall/CRC. doi: 10.1201/9781315372495
- Miller, J. (2000). Applying meta-analytical procedures to software engineering experiments. *J. Syst. Softw.* 54, 29–39. doi: 10.1016/S0164-1212(00)00024-8
- Miller, J. (2005). Replicating software engineering experiments: a poisoned chalice or the holy grail. *Inf. Softw. Technol.* 47, 233–244. doi: 10.1016/j.infsof.2004.08.005
- Nakagawa, S., and Schielzeth, H. (2013). A general and simple method for obtaining R² from generalized linear mixed-effects models. *Methods Ecol. Evol.* 4, 133–142. doi: 10.1111/j.2041-210X.2012.00261.x
- Namin, A. S., and Andrews, J. H. (2009). "The influence of size and coverage on test suite effectiveness," in *Proceedings of the eighteenth international symposium on Software testing and analysis* (Chicago, IL), 57–68. doi: 10.1145/1572272.1572280
- Navarro, D. J. (2019). Between the devil and the deep blue sea: tensions between scientific judgement and statistical model selection. *Comput. Brain Behav.* 2, 28–34. doi: 10.1007/s42113-018-0019-z
- Nosek, B. A., and Errington, T. M. (2020). What is replication? *PLoS Biol.* 18:e3000691. doi: 10.1371/journal.pbio.3000691
- Pearl, J. (2010). "Causal inference," in *Causality: objectives and assessment* (Whistler, BC), 39–58. doi: 10.1017/CBO9780511803161
- Pfleeger, S. L. (1999). Albert einstein and empirical software engineering. *Computer* 32, 32–38. doi: 10.1109/2.796106
- Pickard, L. M., Kitchenham, B. A., and Jones, P. W. (1998). Combining empirical results in software engineering. *Inf. Softw. Technol.* 40, 811–821. doi: 10.1016/S0950-5849(98)00101-3
- Rafique, Y., and Mišić, V. B. (2012). The effects of test-driven development on external quality and productivity: a meta-analysis. *IEEE Trans. Softw. Eng.* 39, 835–856. doi: 10.1109/TSE.2012.28
- Ralph, P. (2018). Toward methodological guidelines for process theories and taxonomies in software engineering. *IEEE Trans. Softw. Eng.* 45, 712–735. doi: 10.1109/TSE.2018.2796554
- Ralph, P., and Tempero, E. (2018). "Construct validity in software engineering research and software metrics," in *Proceedings of the 22nd international conference on evaluation and assessment in software engineering* (Christchurch), 2018, 13–23. doi: 10.1145/3210459.3210461
- Riley, R. D., Lambert, P. C., and Abo-Zaid, G. (2010). Meta-analysis of individual participant data: rationale, conduct, and reporting. *BMJ* 340:c221. doi: 10.1136/bmj.c221
- Rosenthal, R., and DiMatteo, M. R. (2001). Meta-analysis: recent developments in quantitative methods for literature reviews. *Annu. Rev. Psychol.* 52, 59–82. doi: 10.1146/annurev.psych.52.1.59
- Santos, A., Gómez, O., and Juristo, N. (2018). Analyzing families of experiments in se: a systematic mapping study. *IEEE Trans. Softw. Eng.* 46, 566–583. doi: 10.1109/TSE.2018.2864633
- Santos, A., Vegas, S., Oivo, M., and Juristo, N. (2019). A procedure and guidelines for analyzing groups of software engineering replications. *IEEE Trans. Softw. Eng.* 47, 1742–1763. doi: 10.1109/TSE.2019.2935720
- Shepperd, M. (2013). "Combining evidence and meta-analysis in software engineering," in *Software engineering: international summer schools, ISSSE 2009-2011. Revised Tutorial Lectures* (Salerno), 46–70. doi: 10.1007/978-3-642-36054-1_2
- Shepperd, M., Ajenka, N., and Counsell, S. (2018). The role and value of replication in empirical software engineering results. *Inf. Softw. Technol.* 99, 120–132. doi: 10.1016/j.infsof.2018.01.006
- Siebert, J. (2023). Applications of statistical causal inference in software engineering. *Inf. Softw. Technol.* 159:107198. doi: 10.1016/j.infsof.2023.107198
- Sjöberg, D. I., and Bergersen, G. R. (2022). Construct validity in software engineering. *IEEE Trans. Softw. Eng.* 49, 1374–1396. doi: 10.1109/TSE.2022.3176725
- Sutton, A. J., and Higgins, J. P. (2008). Recent developments in meta-analysis. *Stat. Med.* 27, 625–650. doi: 10.1002/sim.2934
- Vaida, F., and Blanchard, S. (2005). Conditional akaike information for mixed effects models. *Biometrika* 92, 351–370. doi: 10.1093/biomet/92.2.351
- Vegas, S., Apa, C., and Juristo, N. (2015). Crossover designs in software engineering experiments: benefits and perils. *IEEE Trans. Softw. Eng.* 42, 120–135. doi: 10.1109/TSE.2015.2467378
- Vehtari, A., Gelman, A., and Gabry, J. (2017). Practical bayesian model evaluation using leave-one-out cross-validation and waic. *Stat. Comput.* 27, 1413–1432. doi: 10.1007/s11222-016-9696-4
- Wagner, S., Mendez, D., Felderer, M., Graziotin, D., and Kalinowski, M. (2020). "Challenges in survey research," in *Contemporary Empirical Methods in Software Engineering*, eds. M. Felderer, and G. H. Travassos (Cham: Springer), 93–125. doi: 10.1007/978-3-030-32489-6_4
- West, B. T., Welch, K. B., and Galecki, A. T. (2022). *Linear Mixed Models: A Practical Guide Using Statistical Software*. Boca Raton, FL: Chapman and Hall/CRC. doi: 10.1201/9781003181064
- Wohlin, C. (2013). "An evidence profile for software engineering research and practice," in *Perspectives on the Future of Software Engineering: Essays in Honor of Dieter Rombach*, eds. J. Münch, and K. Schmid (Berlin: Springer), 145–157. doi: 10.1007/978-3-642-37395-4_10
- Wohlin, C., Runeson, P., Höst, M., Ohlsson, M. C., Regnell, B., Wesslén, A., et al. (2012). *Experimentation in Software Engineering*, Vol. 236. Berlin: Springer. doi: 10.1007/978-3-642-29044-2
- Yang, B. (2002). Meta-analysis research and theory building. *Adv. Dev. Hum. Resour.* 4, 296–316. doi: 10.1177/1523422302043005
- Zain, Z. M., Sakri, S., and Ismail, N. H. A. (2023). Application of deep learning in software defect prediction: systematic literature review and meta-analysis. *Inf. Softw. Technol.* 158:107175. doi: 10.1016/j.infsof.2023.107175