Chapter 10

Threats to Validity in Empirical Software Security Research

Daniela S. Cruzes and Lotfi ben Othmane

CONTENTS

10.1 Introduction

Empirical research in secure software engineering is increasingly important to advancing the state of the art in a scientific manner [16, 17]. Several recent results have pointed to problems related to how security research is conducted or reported in a way that is not advancing the area scientifically. Science of Security (SoS) is an area of research that seeks to apply a scientific approach to the study and design of secure and trustworthy information systems [16, 17]. The core purpose of science is to develop fundamental laws that let us make accurate predictions. Currently, the only prediction we can usually make confidently in secure software engineering is that a system will eventually fail when faced with sufficiently motivated attackers. However, there is a need and an opportunity to develop fundamental research to guide the development and understand the security and robustness of the complex systems on which we depend.

Secure software engineering research is a long way from establishing a scientific approach based on the understanding of empirical evaluation and theoretical foundations as developed in other sciences, and even from software engineering in general. Many of our security and privacy best practices are derived from anecdote, not from careful, evidence-based research [23]. The area suffers from a lack of credible empirical evaluation, a split between industry practice and academic research, and a huge number of methods and method variants, with differences little understood and artificially magnified [17]. There is little empirical evidence on how to implement security practices in the software industry [35]. For example, in 2010, Alnatheer et al. found 62 papers on the topics “agile” and ’security’; of these, only five were empirical [1].

The critical element of any empirical study is to analyze and mitigate threats to the validity of the results. A number of summaries, approaches, and lists of validity threats have been presented in the literature of other areas to help a researcher analyze validity and mitigate threats in different domains [12, 8, 36]. However, they are rarely used in secure software engineering research. It may not be clear to a researcher in the field whether the existing checklists are consistent or which one applies to a particular type of study.

This chapter discusses how validity threats can be analyzed and mitigated in secure software engineering research studies. It first defines validity and threats to validity in Section 10.2. Next, it discusses threats to validity for quantitative research in Section 10.3 and threats to validity for qualitative research in Section 10.4. The chapter includes examples that demonstrate how authors have discussed and addressed threats to validity in secure software engineering research. Section 10.5 concludes the chapter.

10.2 Defining Validity

The validity of a study is the extent to which its design and conduct are likely to prevent systematic errors, or bias [13]. This is distinct from the larger question of how a piece of research might have low quality, since quality has more aspects than validity. In non-technical terms, validity is concerned with "How might the results be wrong?," not with the larger question of "How might this research be bad?," although they do often overlap. Validity is a goal, not something that can be proven. However, in some specific settings, it is possible to form a procedure to ensure the validity of the study—similar to ensuring that a software program is secure.

A *validity threat* is a specific way in which a study might be wrong [18]. The analysis of threats to validity has become a common practice in empirical software engineering studies. For instance, 54% of papers published in ICSE (2012,2013), FSE (2011,2013), and EMSE (2011 to 2013) discussed threats to validity of the study they described in some way [34]. For quantitative research in software engineering, such as experiments, specific advice on validity analysis and threats was given by Wohlin et al. [37], structured according to previous works of Campbell and Stanley (1963) [8] and revised by Cook and Campbell (1979) [10], where they define four main categories of threats, namely: conclusion, internal, construct, and external. There are no software engineering-specific results in qualitative research methods and we have to turn to other fields of study [27, 22, 33]. Analysis of threats to validity allows us to incorporate rigor and subjectivity as well as creativity into the scientific process.

A fundamental difference between the two resides in the fact that the quantitative research paradigm has historically been closely linked to positivism, while qualitative research has to incorporate rigor and subjectivity as well as creativity into the scientific process. Epistemologically, qualitative research stretches between positivism as one extreme and interpretivism as the other [32, 9]. Positivists view humans as a data source like any other that can be sensed, measured, and positively verified [32]. Positivism involves a definite view of scientists as analysts or interpreters of their subject matter, while in the interpretive paradigm, the central endeavor is to understand the subjective world of human experience, concentrating upon the ways in which people view their social world [9]. In the interpretive approach, efforts are made to get inside the person's mind and to understand from within in order to retain the integrity of the phenomenon being investigated. This approach resists the imposition of external form and structure from the positivistic approach, since this reflects the viewpoint of the observer as opposed to that of a directly involved actor [32, 9]. Both positivistic and interpretivistic scientists are interested in assessing whether they are observing, measuring or identifying what they think and say they are, and that is why there is a need to be concerned about the possible threats to validity in any type of study. However, as their research questions, methods, and views on reality differ, so do the methods to assess the quality of their work [32].

Threats to validity will always be present in any empirical research. The goal is to try to mitigate as many known possible threats to research validity as possible. However, the analysis of threats to validity is often considered to be a post-research walk-

through of limitations with limited actual effect on the study [18]. Instead, threats to validity should be considered during the analysis and design of the research, and in reporting the results. Researchers should use techniques to mitigate the threats when possible and needed.

10.3 Validity for Quantitative Research

Quantitative research uses statistical methods to answer questions using data collected from phenomena or participants. The goal of a quantitative research study is to evaluate a claim. This requires the design of treatments (a treatment is a method or process that deals with something or someone) and observe their effects or outputs. The common methods used to investigate secure software engineering challenges are: experimentation, questionnaires and surveys, and data analytics. The research results obtained from these methods are incomplete without analyzing the threats to the validity of the study, e.g., is the size of the sample sufficient? The threats to validity limit the scope of the research and reduce the applicability of the research [24]. For example, a study performed in one company applies to only that company. Nevertheless, the results of a given study should be valid for the population from which the sample is drawn, i.e., adequately valid [37].

In this section, we discuss the categories of threats to validity that are commonly used in quantitative research and the main threats to validity that apply to quantitative research methods.

10.3.1 Categories of threats to validity

There are different classification schemes of threats to validity [37]. Two categorizations are the most common. According to Campbell and Stanley [8], threats to validity are either internal or external. Internal validity concerns "controlling" the aspects of the experiment's setting to ensure that the outcomes are caused only by the introduced technique(s) [34]. External validity refers to showing a real-world effect, but without knowing which factors actually caused the observed difference [34]. Cook and Campbell extended the list of validity categories to: conclusion, internal, construct, and external validity [10]. The latter classification has been adopted in software engineering [37, 24]. In the following, we describe the four validity categories.

10.3.1.1 Conclusion validity

Every empirical study establishes relationships between the treatment, represented by the independent variables, and the outcomes, represented by the dependent variables. The researcher derives conclusions from these relationships, which should have practical use. Conclusion validity refers to the belief in the ability to derive conclusions from the relationships. Threats to conclusion validity are limitations to

the study that affect the ability to derive conclusions about the relations between the independent variables and the dependent variable [37].

10.3.1.2 Internal validity

This validity refers to the belief that the changes to the dependent variable A are solely caused by changes of the independent variable set S of the model. Threats to internal validity are influences that can affect the independent variables with respect to causality [37]. The conditions to claim internal validity are [24]:

1. Variable A is related to variable set S;
2. The direction of the relationship is known;
3. The set S is complete, that is, the relationship between variable A and variable set S is not caused by "other" variables.

10.3.1.3 Construct validity

Empirical research is usually performed to check theoretical concepts with respect to a specific phenomenon. Construct validity refers to the belief that the dependent variables and independent variables represent the theoretical concept of the phenomenon accurately. Threats to construct validity are, in general, related to the design of the study or to social factors [37].

10.3.1.4 External validity

Empirical studies are usually performed in the context of specific settings; a study would be performed on a set of software, on a selected set of subjects/participants, etc. External validity refers to the generalization of the results, e.g., it being "safe" to apply the results of a software study to all software of that type. Threats to external validity are conditions that limit the ability to generalize the study results.

10.3.2 Taxonomy of threats to validity

Wohlin et al. [37] developed a threats to validity list inspired by the lists of Cook and Campbell [10]. The list was extended by Malhotra [24]. Table 10.1, Table 10.2, Table 10.3, and Table 10.4 describe respectively the main conclusion, internal, construct, and external validity threats that we believe should be considered in quantitative research. The tables are inspired by the lists of Wohlin et al. [37] and Malhotra [24] and show the research method(s) that each threat applies to and (when possible) provide example(s) of (secure software engineering) publications that considered that threat.

We use "EXP" for experimentation, "QS" for questionnaires and surveys, "DA" for data analytics, and "All" for all three methods.

Table 10.1: Threats to conclusion validity in quantitative research.

Threat	Description	Method	Examples
Statistical validity	Statistical tests have confidence and power, which indicate the ability of the test to assert a true pattern. Low confidence (or power) implies that the results are not conclusive and don't permit deriving conclusions.	All	[19, 11, 30]
Statistical assumptions	Some statistical tests and methods (e.g., prediction and forecasting) use assumptions, such as normality and independence of the data, or independence of the variables. Violations or absence of tests for the assumptions for a given test/method threaten the ability to use the given test/algorithm.	All	
Lack of expert evaluation	Interpreting the results often requires having deep knowledge about the context of the collected data. The results may also include critical hidden facts, which only experts can point out.	All	[7, 5]
Fishing for the result	Fishing for specific results (often results that conform to the researcher hypotheses) impacts the study setup and design. The researcher could "unintentionally" draw conclusions that are not correct for the study setup and design.	All	
Reliability of the measures	Measurements of independent variables should be reliable: measuring the concept twice should provide the same result. Questionnaire wording is an example of causes of this threat.	All	
Reliability of treatment implementation	The implementation of the treatment should follow a standard and it should be the same for all subjects.	All	[29]
Lack of data preprocessing	The quality of raw data is often not excellent. Researchers need to explore them to identify problems, such as missing data, outliers, and wrong data values, e.g., values that do not follow the codification rules.	All	

Table 10.2: Threats to internal validity in quantitative research.

Threat	**Description**	**Method**	**Examples**
Deficiency of treatment setup	The treatment setup is sometimes not appropriate, which may impact the results. For example, noise and tool performance could impact the results of a study, when they are not related to the treatment of the study.	All	[11]
Ignoring relevant factors	Factors not considered in the experiment setup sometimes impact the study results, such as the usability of the tools used in the experiment and their performance.	All	[29]
History	A study composed of a set of treatments applied at different occasions may be impacted by the history threat. Treatments may be given to the same object at several occasions, each of which is associated with specific circumstances, such as time and location. The change in circumstances may impact the results.	QS	
Maturation	The subjects may react differently as time passes while they perform the treatment: some may become bored and others may become motivated.	QS	
Testing	The subjects may behave differently towards the treatment if they do it several times: they may learn the results and adapt their responses accordingly.	QS	
Treatment design	The artifacts used in the treatment. such as the data collection form and the documents used as information source, could affect the results if not well designed and tested.	All	
Subject selection	Subjects for studies are selected to represent a population. The selection method affects the results and their interpretation. The group of subjects that participate in a study is always heterogeneous. The difference between individuals should not be the dominant factor for the study results: the treatment should be the dominant factor.	QS	[19]

Sample selection	Data are usually collected from data sources that represent the context of the study, such as NVD database,[1] or open source logs and artifacts. The data sample should be representative of the studied type of data.	DA	[7, 2]
Incompleteness of data	Researchers often use heuristics or keyword searches to select records from data sources that represent the data required for the given study. These techniques may fail to identify all the expected records from the data sources.	DA	[7, 2, 29]
Mortality	Some of the subjects selected for a given treatment may drop out of the treatment. This should be considered when evaluating the impact of the given treatment on the subjects. Drop-out subjects should be removed from the treatment.	QS	[31]
Imitation of treatment	This applies to studies that require different subjects/groups to apply different methods/techniques and use the responses to compare the methods and techniques. The subjects/groups may provide responses influenced by their experience and knowledge about the evaluated methods if they learn that these methods/techniques are being applied by other subjects/groups.	QS	
Motivation	A subject may be motivated or resistant to use a new approach/method/technique. This may affect their response/performance in applying either the old or the new approach/method/technique.	EXP, QS	

Table 10.3: Threats to construct validity in quantitative research.

Threat	**Description**	**Method**	**Examples**
Theory definition	The measured variables may not actually measure the conceptual variable. An experiment derived from an insufficiently defined theory does not represent the theory. For example, comparing two methods requires that both use the same metrics for measuring the given variables.	All	[11]

[1] https://nvd.nist.gov/

Mono-operation bias	The study should include more than one independent variable, one treatment, and one subject. Discovering a phenomenon from one variable, case, or subject implies that a theory may exist but may not confirm the theory.	All	
Appropriateness of data	Researchers often use heuristics or keyword searches to select records from data sources. These techniques may result in the extraction of records that are not related to the given study.	All	[7]
Experimenter bias	This happens when a researcher classifies artifacts /data based on his/her own perception or understanding rather than an objective metric. The perception may not be correct.	EXP, DA	[7, 2]
Mono-method bias	Using only one metric to measure a variable results in a measurement bias that can mislead the experiment. For example, using only file-size to measure software complexity could be misleading.	All	[19]
Measurement metrics	The measurement method and the details of the measurement impact the study results. For example, the number of years of experience in security may impact the time it takes to fix security vulnerabilities, while having experience may or may not have much impact.	All	[7, 19, 2]
Interaction with different treatments	A subject that participates in a set of treatments may provide biased responses; his/her responses could be impacted by the interactions of the treatments of the study.	QS	
Treatment testing	A study construction needs to be tested for quality assurance. However, the responses of subjects participating in the study test are affected by their experience with the treatment.	QS	
Hypothesis guessing	Some subjects try to figure out the intended outcomes of studies they are involved in and adapt their responses based on their guesses.	QS	
Evaluation apprehension	Subjects may behave in a different way when evaluated, e.g., review their code more thoroughly. This impacts the truth of the evaluated responses.	QS	

Experimenter expectations	The subjects may have expectations of the experiment and may provide answers accordingly. The study should formulate the treatment to mitigate that, such as asking the questions in different ways.	QS

Table 10.4: Threats to external validity in quantitative research.

Threat	Description	Method	Examples
Representation of the population	The selected subjects/groups should represent the population that the study applies to. For example, security experts cannot represent software developers in a study that investigates a secure software development aspect.	All	[7, 2]
Representation of the setting	The setting of the study should be representative of the study goal. For example, tools used in the study should represent a real setting–not old ones.	All	[7, 2]
Context of the study	The time and location of the study impacts the ability to generalize its results. For example, a study performed on use of code analysis tools only in Germany should not be generalized; developers in other countries may have a different awareness level with respect to code analysis.	All	[19]

10.3.3 Examples of threats to validity analysis

This section discusses three examples of threats to validity analysis in three publications: a questionnaire- (or survey-) based [4] study, an experiment-based study [7], and a data analytics-based study [5]. The examples are informative; the analysis could be improved. However, we classified the validity threats discussed in these publications based on the threat taxonomy of Subsection 10.3.2.[2]

10.3.3.1 Surveys and questionnaires

This subsection gives an example of threats to validity analysis for a questionnaire-based empirical study, which was reported in [4]. An overview of the study and the threats to validity analysis follow.

Overview of the study. Attacker capability is the ability to access a set of resources

[2]The description is based on our understanding of the threats as discussed in the publications.

of an Information System (IS) to exercise threats. For example, an attacker who wants to ”interrupt” a security camera of a corporation and knows how to push the power off button of the camera or how to cut the communication cable can cause the threat only if they have the capability “physical access to the camera.” The authors hypothesize that security experts are less uncertain about their estimations of threat likelihoods when they consider attacker access capabilities. To answer the question, two questionnaires were sent to a set of security experts to report about their risk estimations for a set of threats to two hypothetical systems, a video conferencing system and a connected vehicles system. The authors compared the uncertainty of experts in evaluating the likelihood of threats considering and not considering attacker capabilities for both system examples. The results of the study suggest that experts are more certain about their estimations of threat likelihoods when they consider attacker capabilities.

Threats to validity. A summary of the threats to the validity analysis of the study follows.

The discussed threats to conclusion validity are:

- Reliability of the measures. The experiment results could be affected by the quality of the questions. The authors addressed this threat by testing the questionnaires before making them available to the participants.
- Statistical assumptions. The sizes of the samples for both examples were checked to be limited. The Student distribution was used in testing the hypothesis.
- Statistical validity. Effect size is used to test whether the difference between two quantities being compared is of practical consequence.

The discussed threats to internal validity are:

- Motivation. The participants were given a video to watch that showed attacks that apply to one of the example systems. This may impact the opinions of the experts.
- History. The experiment results could be affected by the fact that each participant had to take successively the two parts of each questionnaire successively, and the hypothesis compares the data of these parts.
- Subject selection. The authors addressed participants who were supposed to be security experts and gave them an authentication code to access the questionnaires. Since the questionnaire was anonymous (as mandated by the data collection regulations), it was not possible to ensure the authenticity of the data.

The threats to construct validity discussed are:

- Theory definition. The authors used a set of independent variables in the experiments that are commonly used for estimating the likelihood of threats, but their effectiveness was not assessed beforehand.
- Evaluation apprehension. There is a potential for difference between perception and reality in questionnaires [21].
- Mono-operation bias. The authors used two examples of systems for the study.

The threats to external validity discussed are:

- Representation of the population. The authors tested the hypothesis using two examples of typical systems.

Observation. We observe that the authors discussed only a set of validity threats. For example, they did not discuss treatment testing and context of the study in the threats to validity section. However, we observe that the authors addressed the threat treatment testing because the questionnaires were tested before making them available for the participants. Not discussing specific validity threats limits the trust in the study, even if they are addressed.

Mitigating threats to validity often impacts the design of the study. For example, we observe that the authors addressed the validity threat mono-operation bias by applying the treatment on two system examples. This improves the validity of the study results. However, it is not always possible to address all the threats in one study. Studies can complement each other.

10.3.3.2 Experiments

This subsection gives an example of threats to validity analysis for experiment-based empirical studies, as reported in [7]. An overview of the study and the threats to validity analysis follow.

Overview of the study. The authors analyzed the security vulnerabilities that could be discovered by code review, identified a set of characteristics of vulnerable code changes, and identified the characteristics of developers that are more likely to introduce vulnerabilities. They analyzed 267,046 code review requests from 10 Open Source Software (OSS) projects and identified 413 Vulnerable Code Changes (VCC). They found that code review can identify the common types of vulnerabilities; the less experienced contributors' changes were 1.8 to 24 times more likely to be vulnerable; the likelihood of a vulnerability increases with the number of lines changed; and modified files are more likely to contain vulnerabilities than new files.

Threats to validity. A summary of the threats to the validity analysis of the study follows.

The discussed threats to conclusion validity are:

- Statistical validity. The dataset of 413 VCCs was built from 267,046 review

requests mined from 10 diverse projects, which is large enough to draw a conclusion with a 95% confidence level.

- Statistical assumptions. The data were tested for normality prior to conducting statistical analyses and used appropriate tests based on the results of the normality test.

The threats to internal validity discussed are:

- Treatment design. The authors selected only projects that practice modern code review supported by Gerrit.[3] The authors believe that using other code review tools should provide the same results because all code review tools support the same basic purpose.
- Sample selection. The authors included most of the public projects managed using the Gerrit tool that contain a large number of code review requests. These projects cover multiple languages and application domains. The authors acknowledge that some of the analyzed projects may not provide a good representation of the types of analyzed security vulnerabilities.
- Incompleteness of data. The authors included data only from projects that practice code review supported by Gerrit. Projects that use other tools were not considered. In addition, the authors excluded a small number of very large code changes under the assumption that they were not reviewed.
- Ignoring relevant factors. OSS projects vary on characteristics like product, participant types, community structure, and governance. This limits the ability to draw general conclusions about all OSS projects from only this single study.

The threats to construct validity discussed are:

- Appropriateness of data. The keyword set used in the study may be incomplete; thus, the search could have missed some data. The authors mitigated this by manually reviewing 400 randomly selected requests. They found only one security vulnerability, which increases the confidence in the validity of the keyword set. In addition, the authors reviewed the comments of review requests that contained at least one keyword and excluded 88% of the review requests. The exclusion was not based on a detailed review of the requests but rather on having the agreement of two reviewers.
- Experimenter bias. Two authors independently inspected and classified each of the 1,348 code review requests to avoid experimenter bias. The authors discussed disagreements and consulted with a third person to address disagreements.

[3] https://www.gerritcodereview.com/

- Measurement method. The study used the number of prior code changes or reviews as a metric of developer experience. The variable is complex and using different measurement methods (e.g., years of experience) could produce different results.

The threats to external validity discussed are:

- Representation of the population. The chosen projects include OSS that vary across domains, languages, age, and governance. Therefore, the results are believed to apply to other OSS projects.

Observation. We observe that the authors took measures to address many of the threats, such as appropriateness of the data and experimenter bias. We also observe that the authors explicitly discussed the validity threats "statistical assumptions" and "ignoring relevant factors"; both are rarely discussed.

10.3.3.3 Security data analytics

This subsection gives an example of threats to validity analysis for a data analytics-based empirical study, which was reported in [5]. An overview of the study and the threats to validity analysis follow.

Overview of the study. The paper is a quantitative investigation of the major factors that impact the time it takes to fix a given security issue based on data collected automatically within SAP's secure development process. The authors used three machine-learning methods to predict the time needed to fix issues and evaluated the predictive power of the prediction models. They found that the models indicate that the vulnerability type has less dominant impact on issue fix time than previously believed and that the time it takes to fix an issue seems much more related to the component in which the potential vulnerability resides, the project related to the issue, the development groups that address the issue, and the closeness of the software release date. The results indicate that the software structure, the fixing processes, and the development groups are the dominant factors that impact the time needed to address security issues.

Threats to validity. A summary of the threats to the validity analysis of the study follows.

The threats to conclusion validity discussed are:

- Statistical validity. The sizes of the data sets were large enough to draw conclusions.
- Reliability of measures. The data is generated automatically and does not include subjective opinions, except for one variable, which is estimated by the experts.

The threats to internal validity discussed are:

- Ignoring relevant factors. There is a consensus in the community that there are many "random" factors involved in software development that may impact the results of data analytics experiments [6]. This applies to this study.
- Deficiency of treatment setup. The data was collected over 5 years. During that time, SAP refined and enhanced its secure software development processes. It was not possible to identify the major process changes along with the times of changes. This could bias the results.

The threats to construct validity discussed are:

- Theory definition. The conclusions are based on the data that SAP collects about fixing vulnerabilities in its software. Changes to the data-collection processes, such as changes to the attributes of the collected data, could impact the predictions and the viability of producing predictions in the first place.
- Mono-operation bias. The authors used three regression methods: Linear Regression (LR), Recursive PARTitioning (RPART), and Neural Network Regression (NNR). However, they did not run the experiment using other single and ensemble regression methods that may apply.

The threats to external validity discussed are:

- Representation of the population. The development teams at SAP develop different types of software, adopt different internal development processes, use different programming languages and platforms, and are located in different cities and countries.

Observation. We observe that the authors explicitly discussed the validity threats "deficiency of treatment setup" and "ignoring relevant factors," which are rarely discussed.

10.4 Threats to Validity for Qualitative Research

Validity in qualitative research has a different set of characteristics than in quantitative studies. The view that methods could guarantee validity was a characteristic of early forms of positivism, which held that scientific knowledge could ultimately be reduced to a logical system that was securely grounded in irrefutable sense data [27]. The validity of qualitative studies depends on the relationship of the conclusions with reality, and no method can fully ensure that the relationship is captured. Although methods and procedures do not guarantee validity, they are nonetheless essential to the process of ruling out validity threats and increasing the credibility of the research conclusions.

Maxwell [27] also affirms that validity is relative; it has to be assessed in relation to the purposes and circumstances of the research, rather than being a context-independent property of methods or conclusions, and methods are only a way of

getting evidence that can help the researcher to rule out these threats. Validity, as a component of the research design, consists of the conceptualization of the possible threats to validity and the strategies used to discover whether they are plausible in the actual research situation, and to deal with them if they are plausible [27].

The proliferation of qualitative research in the past several decades has advanced the science of diverse areas of software engineering, but not much debate has ensued regarding epistemological, philosophical, and methodological issues of these studies in our area yet.

Lincoln and Guba [22] were among the first to start redefining threats to validity concepts to suit qualitative research. They substituted reliability and validity with the parallel concept of "trustworthiness," consisting of four aspects, credibility, transferability, dependability, and confirmability, with credibility as an analog to internal validity, transferability as an analog to external validity, dependability as an analog to reliability, and confirmability as an analog to objectivity. They recommended the use of specific strategies to attain trustworthiness such as negative cases, peer debriefing, prolonged engagement and persistent observation, audit trails and member checks (see next section). Also important were characteristics of the investigator, who must be responsive and adaptable to changing circumstances, holistic, having processional immediacy, sensitivity, and ability for clarification and summarization [22]. These authors were followed by others who either used Guba and Lincolns' criteria or suggested different labels to meet similar goals or criteria. This resulted in a plethora of terms and criteria introduced for minute variations and situations in which rigor could be applied, as shown in Table 10.5.

Table 10.5: Validity criteria from different authors (adapted from Whittemore et al. [36])

Authors	Validity Criteria
Lincoln and Guba (1985) [22]	Credibility, transferability, dependability and confirmability
Sandelowski (1986) [33]	Credibility, fittingness, auditability, confirmability, creativity, artfulness
Maxwell (1992, 1996) [25, 26]	Descriptive validity, interpretive validity, theoretical validity, evaluative validity, generalizability
Eisenhart and Howe (1992) [15]	Completeness, appropriateness, comprehensiveness, credibility, significance
Leininger (1994) [20]	Credibility, confirmability, meaning in context, recurrent patterning, saturation, transferability

We adopt a conservative approach and propose the use of the definitions provided by Lincoln and Guba [22] as described below.

- Credibility: This is the quality of being convincing or believable, worthy of trust. Lincoln and Guba say that it is a twofold task. First, to carry out the inquiry in such a way that the probability that the findings will be found to be credible is enhanced. Second, to demonstrate the credibility of the findings by having them approved by the constructors (participants) of the multiple realities being studied.

- Transferability: Refers to the degree to which the results of the qualitative research can be generalized or transferred to other contexts or settings. It depends on the degree of similarity between sending and receiving contexts. Therefore, transferability inferences cannot be made by an investigator who knows only the sending context. The best advice to give to anyone seeking to make a transfer is to accumulate empirical evidence about contextual similarity; the responsibility of the original investigator ends in providing sufficient descriptive data to make such similarity judgments possible.

- Dependability: Refers to stability and reliability of data over time and conditions. Demonstrating credibility is one way to demonstrate dependability. Lincoln and Guba also point to triangulation, and replication as the means to establish dependability.

- Confirmability: Refers to neutrality; that is, findings must reflect the participants' voice and conditions of the inquiry, and NOT the researcher's bias, perspective, or motivations. The main method proposed by Lincoln and Guba is the confirmability audit, keeping referential adequacy of the data, triangulation, keeping a reflexive journal and raw data (including electronically recorded materials, written notes, etc.), and process notes including methodological notes (procedures, designs, strategies, rationale).

10.4.1 Techniques for Demonstrating Validity in Qualitative Studies

Several techniques contribute to validity in qualitative research, such as the methods employed in differing investigations to demonstrate or assure specific validity criteria [36]. Qualitative research methodology requires a multitude of strategic choices, many of which are practical; however, the rationale for inquiry is not based on a set of deterministic rules. Contextual factors contribute to the decision as to which technique will optimally reflect specific criteria of validity in particular research situations. Techniques can be variously employed, adapted, and combined to achieve different purposes.

Whittemore et al. [36] divide these techniques into four main categories: design consideration, data generation, analytics and presentation. We combined the techniques from Whittemore et al. [36] with the ones from Maxwell [27] and Lincoln and Guba [22]. Table 10.6 shows the techniques that we believe are the most relevant to secure software engineering research.

Table 10.6: Techniques for addressing threats to validity in qualitative research.

Type of Technique	Technique
Design Consideration	Developing a self-conscious research design
	Sampling decisions (i.e., sampling adequacy)
	Employing triangulation
	Peer debriefing
	Performing a literature review
	Sharing perquisites of privilege
Data Generation	Articulating data collection decisions
	Demonstrating prolonged engagement
	Rich data – demonstrating persistent/intense observation
	Referential adequacy – providing verbatim transcription)
	Reflexive journaling
	Demonstrating saturation
Analytics	Articulating data analysis decisions
	Member checking or respondent validation
	Expert checking
	Exploring rival explanations, discrepant evidence and negative cases
	Triangulation
	Drawing data reduction tables
Presentation	Providing evidence that supports interpretations
	Acknowledging the researcher perspective
	Thick descriptive data

Most of the techniques are self-explanatory, but for some of them it is important to provide a description and further details. The following descriptions are based on Maxwell [27] and Lincoln and Guba [22]:

- Intensive long-term involvement: Lengthy and intensive contact with the phenomena (or respondents) in the field to assess possible sources of distortion and especially to identify saliencies in the situation. It provides more complete data about specific situations and events than any other method. Not only does it provide a larger amount and variety of data, it also enables the researcher to check and confirm the observations and inferences. Repeated observations and interviews, as well as the sustained presence of the researcher in the study setting can help rule out spurious associations and premature theories. It also allows a much greater opportunity to develop and test alternative hypotheses in the course of the research process. Finally, the period of prolonged engagement also provides the investigator an opportunity to build trust.

- Rich data / persistent observation: Both long-term involvement and intense interviews enable the researcher to collect rich data, i.e., data that are detailed

and varied enough to provide a full revealing picture of what is going on. The purpose of persistent observation is to identify those characteristics and elements in the situation that are most relevant to the problem or issue being pursued and focused on them in detail. In interview studies, such data generally require verbatim transcripts of the interviews, not just notes of what the researcher felt was significant.

- Respondent validation or member checking: This is about systematically soliciting feedback about the data and conclusions from the people under study. This is of the most important way of ruling out the possibility of misinterpreting the meaning of what participants say and do and the perspective they have on what is going on, as well as being an important way of identifying biases and misunderstandings of what is observed.

- Searching for discrepant evidence and negative cases: This is a key part of the logic of validity testing in qualitative research. Instances cannot be accounted for by a particular interpretation or explanation that can point to important defects in that account. The basic principle here is that the researcher needs to rigorously examine both the supporting and the discrepant data to assess whether it is more plausible to retain or modify the conclusion, being aware of all of the pressures to ignore data not fitting the conclusions.

- Triangulation: Collecting information from a diverse range of individuals and settings, using a variety of methods, and at times, different investigators and theories. This strategy reduces the risk of chance associations of systematic biases due to a specific method, and allows a better assessment of the generality of the explanations of the developers.

- Peer debriefing: Exposing oneself to a disinterested professional peer to ”keep the inquirer honest,” assists in developing working hypotheses, develops and tests the emerging design, and facilitates emotional catharsis.

- Member checking: The process of continuous, informal testing of information by soliciting reactions of respondents to the investigator’s reconstruction of what he or she has been told or otherwise found out and to the constructions offered by other respondents or sources, and a terminal, formal testing of the final case report with a representative sample of stakeholders. Member checking is both informal and formal, and it should occur continuously.

- Thick descriptive data: Narrative description of the context so that judgments about the degree of fit or similarity may be made by others who may wish to apply all or part of the findings elsewhere. (Although it is by no means clear how thick a thick description needs to be.) Dybå et al. [14] discuss how to define what context variables should be accounted for in a study.

- Referential adequacy: A means for establishing the adequacy of critiques written for evaluation purposes under the connoisseurship model. The recorded materials provide a kind of benchmark against which later data analysis and

interpretations (the critiques) can be tested for adequacy. Aside from the obvious value of such materials for demonstrating that different analysts can reach similar conclusions given whatever data categories have emerged, they can also be used to test the validity of the conclusions.

- Reflexive journaling: A kind of diary in which the investigator, on a daily basis or as needed, records a variety of information about themselves (what is happening in terms of their own values and interests and for speculation about growing insights) and method (information about methodological decisions made and the reasons for making them) in addition to the daily schedule and logistics of the study.

In the following subsection, we discuss examples of studies that discuss some of these validity threats.

10.4.2 Examples of Threats to Validity for Qualitative Studies

We performed a search in five systematic reviews in software security and did not find many examples of how researchers handle threats to validity in their studies. The only two qualitative studies we found in these systematic reviews that deal with or mention threats to validity are described as examples below.

10.4.2.1 Case Studies

This subsection gives an example of threats to validity analysis for a questionnaire-based empirical study, which was reported in [28]. An overview of the study and the threats to validity analysis follow.

Overview of the Study. Today, companies are required to have control over their IT assets, and to provide proof of this in the form of independent IT audit reports. However, many companies have outsourced various parts of their IT systems to other companies, which potentially threatens the control they have over their IT assets. To provide proof of having control iver outsourced IT systems, the outsourcing client and outsourcing provider need a written service-level agreement (SLA) that can be audited by an independent party. SLAs for availability and response time are common practice in business, but so far there is no practical method for specifying confidentiality requirements in an SLA. Specifying confidentiality requirements is hard because in contrast to availability and response time, confidentiality incidents cannot be monitored: attackers who breach confidentiality try to do this unobserved by both client and provider. In addition, providers usually do not want to reveal their own infrastructure to the client for monitoring or risk assessment. Elsewhere, the authors have presented an architecture-based method for confidentiality risk assessment in IT outsourcing. The authors adapt this method to confidentiality requirements specification, and present a case study to evaluate this new method. The method is based on specifying confidentiality requirements according to risk assessment results.

Threats to Validity. A summary of the threats to the validity analysis of the study follows. The discussed threats are:

- Credibility: The authors say that: "...To validate a method, we eventually need a realistic context in which the method is applied. Applying it to a toy problem is fine for illustration, and testing in an experiment is good for improving our understanding of the method, but in order to know whether the method will work in practice, it has to be used in practice. This could be done by a field experiment, in which practitioners use the method to solve an experimental problem. This is extremely expensive but not impossible. In our case, we opted for the more realistic option, given our budget, of using the method ourselves for a real world problem."
- Transferability: The authors applied their method for confidentiality risk assessment and comparison twice with similar results, both in multinational industrial companies where confidentiality was not a critical requirement until external regulators enforced it. The authors also state where the transferability of the results may apply: Operating in highly competitive markets, these companies are very cost-sensitive and they will therefore not aim at maximum confidentiality. This might well be different in privacy-sensitive organizations such as health care or insurance companies, or in high confidentiality organizations such as the military. Nevertheless, confidentiality is not the highest-priority requirement for the context of the study. All of this supports reusability to any context that satisfies the three assumptions, with similar answers to the research questions for those contexts.
- Dependability: The authors say: "...We answered the reusability question by identifying the conditions under which the methods can be used, and actually showing that it could be used in another case satisfying these assumptions. Like all inductive conclusions, our conclusion that the method can be used in other cases is uncertain, but because we used analytic reasoning rather than statistical reasoning, we cannot quantify this uncertainty." Thereby, the authors have shown that they are concerned about the reliability of the results. However, the authors affirm that repeatability of the results needs further research.
- Confirmability: The authors says "we find no reasoning errors or observational mistakes so we claim these answers are valid," but the aspect of neutrality is not clear cut when the authors were the ones using the method and providing feedback. The findings still reflect participants' voice and conditions of the inquiry, but it remains unclear to what extent the authors took their own researcher biases, perspective, or motivations into account.

Observations. Even though the authors did not use the nomenclature we used above, they were quite conscious of revealing the possible threats to the study results and what they had done to mitigate possible threats. In addition, the authors were very good to describe the context of the case study, so transferability can be established more easily.

10.4.2.2 Interviews

This subsection gives an example of threats to validity analysis for an interview-based empirical study, which was reported in [3]. An overview of the study and the threats to validity analysis follow.

Overview of the Study. Agile methods are widely employed to develop high-quality software, but theoretical analyses argue that agile methods are inadequate for security-critical projects. However, most agile-developed software today needs to satisfy baseline security requirements, so we need to focus on how to achieve this for typical agile projects. The author provides insights from the practitioner's perspective on security in agile development and reports on exploratory, qualitative findings from 10 interviews. The goal of the study is to expand on the theoretical findings on security-critical agile development through an exploration of the challenges and their mitigation in typical agile development projects.

Threats to Validity. A summary of the threats to the validity analysis of the study follows.

- Credibility: The participants' views were collected in interviews only; the possible threats of the chosen design are not addressed.
- Transferability: The author says: "We conducted the sampling of the participants in a way to represent a wide variety of agile development contexts. Sampling dimensions included the interviewee's process role and project characteristics, such as team size and development platform." This statement refers to the degree to which the results of the qualitative research can be generalized or transferred to other contexts or settings.
- Dependability: Not discussed in the paper. But the author mentions: Since the sample size is limited for interviews, we focused on covering a broad range of development contexts. The results are, by study design, not sound and representative, but extends the prior theoretical findings with a practical perspective and offers a description as an initial hypothesis for further research.
- Confirmability: The author mentions: "We directly contacted the interviewees primarily through agile development meet-ups. The interviews lasted between 30 and 60 minutes. Based on detailed notes on each interview, we structured the statements by common concepts iteratively with each interview and clustered related aspects to derive the findings on challenges and mitigations." Regarding the neutrality aspect, the author mentions: The interviews offer only subjective data and are prone to researcher or participant bias, but does not explain what he did to mitigate some of these threats.

Observations. The author did not specify the threats to the validity of the studies in detail and failed to show how he tried to mitigate some threats to validity.

10.5 Summary and Conclusions

Empirical research publications in software engineering, in general, mention limitations of the reported results, often without naming the threats. The threats to validity are discussed, in general, as descriptive arguments, without compliance with (specific) validity threats taxonomy. This is practiced in both quantitative and in qualitative studies.

We believe that the main reasons for doing so include that (1) the threats that apply to a given study depend on the study purpose, setup, and context–some of the threats provided in taxonomy of validity threats do not apply to a given study; (2) the need of the authors to have a narration and to respect the limitations on the publication size; and (3) the need to keep the researchers aware of the need for describing the threats to validity of their studies and how they have tried to mitigate the possible threats.

There are two implications of this practice. First, information about threats to validity analysis is, in general, incomplete, as the absence of reports on the status of a given threat to validity that applies to a given study implies neither that the threat is addressed (that is, it is not discussed because the answer is implicit) nor that it is not addressed. Second, absence of uniform reporting of threats to validity hinders the possibility of comparing studies and limits the credibility of systematic literature reviews that try to summarize the knowledge acquired through empirical research on a specific topic.

While validity categories are being increasingly adopted, there is no adoption of validity threat checklists, nor common terminology. Use of threat checklists will help to formally evaluate the validity of studies and to advance the knowledge on the given topic.

Acknowledgment

This work was supported by the SoS-Agile (Science of Security in Agile Software Development) project, funded by the Research Council of Norway under the grant 247678/O70 and by the Hessian LOEWE excellence initiative within CASED.

References

[1] A. Alnatheer, A. M. Gravell, and D. Argles. "Agile Security Issues: An Empirical Study." In: *Proc. of the 2010 ACM-IEEE International Symposium on Empirical Software Engineering and Measurement*. ESEM '10. Bolzano-Bozen, Italy, 2010, 58:1–58:1.

[2] S. S. Alqahtani, E. E. Eghan, and J. Rilling. "Tracing known security vulnerabilities in software repositories – A Semantic Web enabled modeling approach." In: *Science of Computer Programming* 121 (2016). Special Issue on Knowledge-based Software Engineering, pp. 153–175.

[3] S. Bartsch. "Practitioners' Perspectives on Security in Agile Development." In: *Proc.Seventh International Conference on Availability, Reliability and Security*. ARES. Prague, Czech Republic, Aug. 2012, pp. 479–484.

[4] L. ben Othmane et al. "Incorporating attacker capabilities in risk estimation and mitigation." In: *Computers & Security* 51 (June 2015). Elsevier, pp. 41–61.

[5] L. ben Othmane et al. "Time for Addressing Software Security Issues: Prediction Models and Impacting Factors." In: *Data Science and Engineering* (Sept. 2016). Springer.

[6] A. Bener et al. "The Art and Science of Analyzing Software Data." In: ed. by C. Bird, T. Menzies, and T. Zimmermann. 1St. Waltham, USA: Elsevier, Aug. 2015. Chap. Lessons Learned from Software Analytics in Practice, pp. 453–489.

[7] Amiangshu Bosu et al. "Identifying the Characteristics of Vulnerable Code Changes: An Empirical Study." In: *Proc. of the 22nd ACM SIGSOFT International Symposium on Foundations of Software Engineering*. FSE 2014. Hong Kong, China, 2014, pp. 257–268.

[8] D.T. Campbell and J.C. Stanley. "Handbook on Research on Teaching." In: Boston, USA: Houghton Mifflin Company, 1963. Chap. Experimental and Quasi-experimental Designs for Research.

[9] L. Cohen, L. Manion, and K. Morrison. *Research Methods in Education*. 7th. Abingdon, Canada: Taylor & Francis, 2011. ISBN: 9781135722036.

[10] T. D. Cook and D. T. Campbell. *Quasi-Experimentation: Design & Analysis Issues for Field Settings*. Boston, USA: Houghton Mifflin, 1979.

[11] M. D'Ambros, M. Lanza, and R. Robbes. "Evaluating Defect Prediction Approaches: A Benchmark and an Extensive Comparison." In: *Empirical Softw. Engg.* 17.4-5 (Aug. 2012), pp. 531–577.

[12] S. M Downing and T. M Haladyna. "Validity threats: overcoming interference with proposed interpretations of assessment data." In: *Medical Education* 38.3 (Mar. 2004), pp. 327–33.

[13] T. Dybå and T. Dingsøyr. "Strength of Evidence in Systematic Reviews in Software Engineering." In: *Proc. of the Second ACM-IEEE International Symposium on Empirical Software Engineering and Measurement*. ESEM '08. Kaiserslautern, Germany, Oct. 2008, pp. 178–187.

[14] T. Dybå, D. I.K. Sjøberg, and D. S. Cruzes. "What Works for Whom, Where, when, and Why?: On the Role of Context in Empirical Software Engineering." In: *Proc. of the ACM-IEEE International Symposium on Empirical Software Engineering and Measurement*. ESEM '12. Lund, Sweden, 2012, pp. 19–28.

[15] M. A. Eisenhart and K. R. Howe. "The handbook of qualitative research in education." In: ed. by M. D. LeCompte, W. L. Millroy, and J. Preissle. San Diego, USA: Academic Press, 1992. Chap. Validity in educational research, pp. 643–680.

[16] D. Evans. *NSF/IARPA/NSA Workshop on the Science of Security - Workshop Report*. Tech. rep. Berkeley, USA: University of Virginia, Nov. 2008.

[17] D. Evans and S. Stolfo. "Guest Editors' Introduction: The Science of Security." In: *IEEE Security Privacy* 9.3 (May 2011), pp. 16–17.

[18] R. Feldt and A. Magazinius. "Validity Threats in Empirical Software Engineering Research - An Initial Survey." In: *Proc. of the Software Engineering and Knowledge Engineering Conference*. Redwood City, CA, USA, 2010, pp. 374–379.

[19] J. S. Giboney et al. "The Security Expertise Assessment Measure (SEAM): Developing a scale for hacker expertise." In: *Computers & Security* 60 (2016), pp. 37–51.

[20] M. Leininger. "Critical issues in qualitative research methods." In: ed. by J. M. Morse. Thousand Oaks, USA: Sage, 1994. Chap. Evaluation criteria and critique of qualitative research studies, pp. 95–115.

[21] R. Likert. "A Technique for the Measurement of Attitudes." In: *Archives of Psychology* 22.140 (June 1932).

[22] Y. S. Lincoln and E. G. Guba. *Naturalistic inquiry*. Beverly Hills, CA: SAGE Publications, Inc, 1985.

[23] L. Madeyski and B. A. Kitchenham. *Reproducible Research - What, Why and How*. Tech. rep. PRE W08/2015/P-020. Wroclaw University of Technology, 2015.

[24] R. Malhotra. "Empirical Research in Software Engineering." In: CRC Press, 2016. Chap. Introductions, pp. 1–31.

[25] J. Maxwell. "Understanding and Validity in Qualitative Research." In: *Harvard Educational Review* 62.3 (Sept. 1992), pp. 279–301.

[26] J. A. Maxwell. *Qualitative research design: An interactive approach*. Thousand Oaks, USA: Sage, 1996.

[27] J. A. Maxwell. *Qualitative research design: An interactive approach*. 3rd Ed. Thousand Oaks, CA: Sage, 2013.

[28] A. Morali and R. Wieringa. "Risk-based Confidentiality Requirements Specification for Outsourced IT Systems." In: *Proc. 18th IEEE International Requirements Engineering Conference*. RE. 2010, pp. 99–208.

[29] P. Morrison et al. "Challenges with Applying Vulnerability Prediction Models." In: *Proc. of the 2015 Symposium and Bootcamp on the Science of Security*. HotSoS '15. Urbana, Illinois, 2015, 4:1–4:9. ISBN: 978-1-4503-3376-4.

[30] I. Myrtveit, E. Stensrud, and M. Shepperd. "Reliability and validity in comparative studies of software prediction models." In: *IEEE Transactions on Software Engineering* 31.5 (May 2005), pp. 380–391.

[31] K. Nayak et al. "Some Vulnerabilities Are Different Than Others." In: *Proc. 17th International Symposium on Research in Attacks, Intrusions and Defenses*. Gothenburg, Sweden, Sept. 2014, pp. 426–446.

[32] C. Robson. *Real World Research*. Second edition. MA, USA; Oxford UK, and Victoria, Australia: Blackwell, 2002.

[33] M. Sandelowski. “The problem of rigor in qualitative research.” In: *Advances in Nursing Science* 8.3 (Apr. 1986), pp. 27–37.

[34] J. Siegmund, N. Siegmund, and S. Apel. “Views on Internal and External Validity in Empirical Software Engineering.” In: *Proc. of the 37th International Conference on Software Engineering - Volume 1*. Florence, Italy, 2015, pp. 9–19.

[35] J. Wäyrynen, M. Bodén, and G. Boström. “Security Engineering and eXtreme Programming: An Impossible Marriage?” In: *Proc. 4th Conference on Extreme Programming and Agile Methods - XP/Agile Universe*. Ed. by Carmen Zannier, Hakan Erdogmus, and Lowell Lindstrom. Calgary, Canada, Aug. 2004, pp. 117–128.

[36] R. Whittemore, S. K. Chase, and C. L. Mandle. “Validity in qualitative research.” In: *Qualitative Health Research* 11.4 (July 2001), pp. 117–132.

[37] C. Wohlin et al. *Experimentation in Software Engineering*. Berlin Heidelberg: Springer-Verlag, 2012.