Challenges and Recommendations to Publishing and Using Credible Evidence in Software Engineering

Claes Wohlin[a], Austen Rainer[b]

[a]*Blekinge Institute of Technology, SE-371 79 Karlskrona, Sweden*
[b]*Queen's University Belfast, 18 Malone Road, Computer Science Building, BT9 5BN, Belfast, Northern Ireland, UK*

Abstract

Context: An evidence–based scientific discipline should produce, consume and disseminate credible evidence. Unfortunately, mistakes are sometimes made, resulting in the production, consumption and dissemination of invalid or otherwise questionable evidence. In the worst cases, such questionable evidence achieves the status of accepted knowledge. There is, therefore, the need to ensure that producers and consumers seek to identify and rectify such situations.
Objectives: To raise awareness of the negative impact of misinterpreting evidence and of propagating that misinterpreted evidence, and to provide guidance on how to improve on the type of issues identified.
Method: We use a case–based approach to present and analyse the production, consumption and dissemination of evidence. The cases are based on the literature and our professional experience. These cases illustrate a range of challenges confronting evidence–based researchers as well as the consequences to research when invalid evidence is not corrected in a timely way.
Results: We use the cases and the challenges to formulate a framework and a set of recommendations to help the community in producing and consuming credible evidence.
Conclusions: We encourage the community to collectively remain alert to the emergence and dissemination of invalid, or otherwise questionable, evidence, and to proactively seek to identify and rectify it.

Keywords: Evidence–based Software Engineering, EBSE, credible evidence, validity, relevance

1. Introduction

An evidence–based scientific discipline should produce, consume and disseminate credible evidence. This puts high requirements on authors, reviewers, editors and readers to ensure that we, as a discipline, produce credible evidence that others can build upon. Unfortunately, mistakes are unintentionally made by producers of evidence and, again unfortunately, peer–reviewers and editors may not find these mistakes. Publications containing questionable evidence are then disseminated into the public domain. Once in the public domain, consumers of such evidence often reuse that evidence without realising they are perpetuating questionable evidence. Like the proverbial Pandora's Box, the questionable evidence becomes extremely difficult to rectify. So in addition to high requirements on producing credible evidence, there are also high requirements on us to ensure we do not disseminate questionable evidence produced by others.

In this paper, we present and discuss cases from literature where the evidence may be challenged, and then of the subsequent dissemination, of evidence that was initially thought to be credible but subsequently shown to be not as credible as originally accepted by the community. Identification of problems based on analysis of cases have been published by others using a similar approach, for example, problems relating to statistical analysis as discussed by Kitchenham *et al.* [1] and [2], as well as Reyes *et al.* [3].

In using actual cases from research we inevitably identify individuals. We emphasise that our focus here is *only* on the collective responsibility of the community of producers and consumers to quality–assure that our evidence is credible.

1.1. The users of evidence

The uses of evidence may be viewed from different perspectives and for different roles. For this paper, we distinguish between the production and the consumption of evidence, and between the researcher and the practitioner. We summarise these distinctions in Table 1.

Given our focus in this paper on evidence–based research, we do not consider in detail the practitioner as a producer of evidence. Where a person working in industry produces evidence following a recognised research process we classify those people as researchers according to Table 1. Practitioners produce information that is used in subsequent research, e.g., documents that are used in grey literature reviews (e.g., [4]), surveys (e.g., [5]), interviews

(e.g., [6]), focus groups (e.g., [7]), and as data in other studies (e.g., [8]). In these cases, the research papers are in most cases published by researchers and hence this situation falls under the category of researchers as producers of the evidence, although the underlying information may come from practitioners. Researchers are interested in the validity of the practitioners' information and may also be interested in the way in which that information was produced e.g., the practitioners' experience. The validity of practitioners' information, and the processes by which such information is produced, are important issues for research to consider, however they lie beyond the scope of the current paper.

Table 1: Perspectives: Producer/consumer and research/practitioner.

Roles	**Producer**	**Consumer**
Researcher	Yes	Yes
Practitioner	N/A	Yes

A researcher may be both a producer of evidence and a consumer of evidence. In the former case, researchers conduct research and produce new evidence. In the latter case, researchers build upon the work produced by others and hence consume research evidence created by other researchers.

1.2. The nature of evidence

Evidence is an extremely difficult concept to properly define. As an example, Schum [9] extensively reviewed a wide range of definitions of the concept of evidence and concluded,"When all is said and done we may not be able to define the word evidence so that everything acceptable in all recognised disciplines is included and everything else is excluded. . . ". Consequently, we present our terminology and definitions later in this paper, briefly contrasting them with prior work.

1.3. Aims, objectives and contributions

There are two aims to this paper: 1) to raise awareness of the negative impact of misinterpreting evidence, either as the producer or as a consumer of that evidence, and of consequently perpetuating that misinterpretation through the producer publishing that evidence and consumers citing the respective publication; and 2) to provide guidance to producers and to consumers on how to reduce the misinterpretation of evidence.

To meet these aims, we formulate three objectives:

1. To identify some of the challenges confronted by producers and consumers when producing, consuming and disseminating evidence.
2. To identify some of the consequences that arise from the production, consumption and dissemination of invalid or otherwise questionable evidence.
3. To provide recommendations to the community on how to address some of these challenges.

The paper contributes a terminology concerning credible evidence, and a set of cases drawn from literature and from our own experiences. The cases highlight some challenges in the production, consumption and dissemination of credible evidence. To address the challenges, a framework for credible evidence is put forward and a set of recommendations for the production, consumption and dissemination of credible evidence are formulated.

1.4. Structure to the paper

The remainder of the paper is organised as follows. In Section 2, we review previous research. Section 3 describes our research approach. In Section 4, we provide the terminology and the definitions for the concepts used for reasoning about the production and consumption of credible evidence. These concepts provide a background for Section 5 where we present several cases, some of them based on literature and some based on our experiences. The cases in Section 5 highlight that there are challenges when producing and consuming credible evidence. In Section 6, we present a framework for credible evidence, together with a set of challenges and recommendations to address these challenges. Section 7 outlines further directions for research and provides a review of the objectives of the paper.

2. Related work

Software engineering researchers use a variety of words and phrases, such as rigour and validity, to refer to credibility and to related concepts. We briefly review relevant previous research in relation to concepts such as credibility, validity, rigour and relevance. We first concentrate on three papers: Prechelt and Petre's [10] discussion of credibility, Ivarsson and Gorschek's [11] development and application of a model for evaluating the rigor and industrial relevance of technology evaluations in software engineering, and Petersen and Gencel's [12] discussion of worldviews in software engineering

research. We then briefly review other research, and then summarise the section. To follow-up on related work, we contrast the concepts presented in the three papers by Prechelt and Petre [10], Ivarsson and Gorschek [11], and Petersen and Gencel [12] respectively with our definitions of credible evidence and its characteristics in Section 4.

2.1. Prechelt and Petre's discussion of credibility

Prechelt and Petre [10] consider the concepts of credibility, validity, relevance and fit–for–purpose. They define credibility as the degree to which you are – or should be – willing to believe the evidence offered and the claims made about it. They recognise validity as a part of credibility and define validity as the extent to which a study and the claims made from it accurately represent the phenomenon of interest. Note that Prechelt and Petre use validity in relation to both the study, as a process of inquiry, and to the claims arising from that study. In addition to validity, another part of credibility is good reporting. A credible study therefore has both high validity and is well reported.

Prechelt and Petre define relevance in terms of the degree to which you are — or ought to be — interested in the evidence and claims. Prechelt and Petre consider it helpful to distinguish credibility from relevance: a relevant study may have limited validity, or poor reporting, and therefore have limited credibility. For fit–for–purpose, Prechelt and Petre [10] assert that different purposes require different standards of evidence, e.g., if you are reviewing previous research to inform a decision you need to make, the greater the impact or consequence of that decision, the more credible the evidence should be to inform that decision.

Prechelt and Petre [10] review a number of research methods. For each method they highlight a sample of credibility issues and relevance issues. In presenting these issues, Prechelt and Petre are complementing the abstract concepts of credibility and relevance with specific and tangible *challenges* that relate to credibility and relevance. For example, Prechelt and Petre identify issues, and therefore challenges, relating to subject bias and task bias.

2.2. Ivarsson and Gorschek's model for evaluating technology evaluations

Ivarsson and Gorschek [11] develop a model for evaluating the industrial relevance and rigour of technology evaluations in software engineering.

They define the rigour of an evaluation in terms of both how the evaluation is performed and how it is reported. There are clear similarities with Prechelt and Petre's [10] work, but differences too. A feature of rigour is that "... the methodology used should be carried out in accordance with corresponding best practices ...". In other words, Ivarsson and Gorschek frame methodological rigour in terms of conformance to best practice for that methodology. Ivarsson and Gorschek recognise validity as a component of rigour. They write, for example, "The validity of the evaluation is discussed in detail where threats are described and measures to limit them are detailed." Again, we distinguish between the rigour of a process of inquiry — or, in Ivarsson and Gorschek's [11] study, an evaluation — and the validity of the information consumed by or produced from that process of inquiry.

For Ivarsson and Gorschek [11], relevance refers to the potential impact the research has on both academia and industry. Ivarsson and Gorschek distinguish academic relevance e.g., the ability to publish papers, and through citations by other researchers – from industrial relevance e.g., the evaluation's value for practitioners seeking to adopt technologies.

Ivarsson and Gorschek [11] also recognise that the model they develop does not consider if the study design is appropriate to address the research questions. In other words, Ivarsson and Gorschek's [11] model does not consider whether the study design is *fit–for–purpose*.

2.3. Petersen and Gencel on worldviews

Petersen and Gencel [12] discuss the relationship between philosophical worldviews, research methods, types of validity and threats to validity. Their work is not explicitly concerned with credibility or with rigour, but the work implicitly recognises that 'weaknesses' in the process of inquiry threaten the validity of information produced by the inquiry. Petersen and Gencel [12] highlight the challenges of defining validity, and therefore of defining credibility, and also highlight the challenges of consistently evaluating the validity of information, information that has been collected and analysed through a variety of different research methods.

2.4. Other research

There are many existing papers that have examined a diverse range of specific aspects related to evidence. For example, Lo *et al.* [13] report responses, from 512 employees, to a survey in which 3,000 Microsoft employees

were asked to rate the relevance of 571 conference papers in software engineering. Respondents were asked to rate the papers as Essential, Worthwhile, Unimportant or Unwise. Lo *et al.*'s [13] study provides a normative result, i.e., examining the degree to which practitioners rated papers as relevant, in contrast to investigating the factors that improve the relevance of a paper. The Lo *et al.* paper is focused on industrial relevance, or possibly even fit-for-purpose since the study is done with employees from one company.

Zhang *et al.* [14] investigate an approach for more effectively searching for and selecting the more relevant papers for a systematic literature review. Given the focus on identifying papers for a systematic literature review, Zhang *et al*'s. [14] paper is primarily targeting academic relevance. Kitchenham *et al.* [15] provide advice on the use of statistical methods in empirical software engineering. In terms of our model in Figure 1, Kitchenham *et al.*'s paper may be understood as seeking to improve the rigour of data analyses so as to improve the validity of the results from that analyses.

Liebchen and Shepperd (first in [16] and then with a follow–up study in [17]) examine the degree to which researchers evaluate the quality of the *data* they used (or reused from others), and the techniques that researchers used to undertake that evaluation and to address (e.g. repair) data quality issues. In terms of our model, Liebchen and Shepperd's work concerns both validity of information and rigour of the processes for improving the validity of information. In their first paper [16], they defined quality in terms of accuracy of data. They recognise that there were other dimensions to quality e.g. timeliness of data, completeness of data, and fit–for–purpose. Liebchen and Shepperd also recognised the importance of meta information for describing the (perceived) quality of the dataset. In the second of their two papers, Liebchen and Shepperd [17] argued for data quality protocols, in particular the value of such protocols for understanding the impact of pre-processing.

2.5. Summary

We have reviewed previous research, concentrating in particular on three studies, and complementing that concentrated focus with a brief and broad recognition of a range of other related research. Clearly there are differences between researchers in the definitions of concepts and in the relationship between several concepts. In the next section, we describe the terminology and definitions we use in relation to credible evidence, and contrast those with the terminology, definitions and models of selected previous research.

3. Research approach

In this section we briefly describe the research approach we used in our analyses.

The research approach we take for our paper is similar in concept to the analytical approach advocated by Yin [18] for case study research, i.e., we do not seek a statistical analyses of a sample (it is, for example, not clear upfront what a meaningful sample would be) but instead analyse 'cases' in relation to a theoretical position. We emphasise that we are not undertaking a case study in the formal way defined by Yin [18] or by Runeson and Höst [19]. This is because we do not have empirical cases of contemporary phenomena in their real–world context. The approach we take has been successfully used by others in previous research, e.g., MacDonell *et al.* [20], Jørgensen and Kitchenham [21] and Kitchenham *et al.* [1].

To identify cases, we began with a 'seed case' that was well–known to the first author of the current paper. This case is henceforth referred to as our *primary case*. This case was first identified when he attended the International Software Metrics Symposium (ISMS) in 1997. One of the papers presented at ISMS highlighted a problem with relating module size and defect density, since they are coupled mathematically due to defect density being equal to the number of defects divided by size. The problem was illustrated with an empirical example taken from a paper published 13 years earlier, in 1984, and written by a pioneer of empirical software engineering. The first author of the current paper has used this problem and empirical example with his students, for more than 20 years, to help demonstrate the challenges of empirical research.

More recently, in 2017, the first author analysed citations to the two papers (i.e., to the paper with the empirical example from 1984, and to the ISMS paper from 1997 that highlighted the problem with the empirical example) comparing the history of citations for each paper. The first author did this to understand whether and how the problem of relating module size and defect density affected subsequent research. The first author effectively conducted a forward snowballing [22] of citations. This forward snowballing identified two additional concerns and these became our second and third cases, henceforth referred to as our *secondary cases*. The first problem is concerned with a paper by Hatton [23], where he states that smaller components are proportionally more unreliable. However, the article by Basili and Perricone [24] does not discuss reliability, and defect density is not the same

as reliability. Second, the citation analysis resulted in identifying a series of papers, where the authors have missed the opportunity to clearly express the synthesised learning from the series of studies.

The primary case and the two secondary cases arising directly from the primary case are discussed in Sections 5.1–5.2. Further details of the analysis related to the primary case are reported in Appendix A.

Having found two new concerns in the analysis, the two authors of the current paper discussed other cases known to them through the academic and grey literature, and from working with industry. These discussions identified five *additional cases*. These subsequent cases, identified through discussion of the literature and of industrial collaborations, are presented in Section 5.3.

Taken together, our collection of cases illustrates some of the challenges with producing credible evidence, challenges that we believe should be communicated to people who intend to consume that (hopefully credible) evidence.

Given the identification of concerns related to credible evidence, we defined a terminology and reference model for credible evidence. We did this to ensure a common vocabulary and understanding, both between the authors and for the potential reader. The terminology and reference model are presented in Section 4 and are based on our review of previous work, which is presented in Section 2. Based on the cases discussed in Section 5, we developed a framework to be used by producers and consumers of credible evidence. This is presented in Section 6.1. Furthermore, to increase both the ability to produce and to consume credible evidence, we developed recommendations. These are presented in Sections 6.2–6.4.

4. Terminology and definitions

We recognised in Section 1 that the concept of evidence is extremely difficult to define. Related concepts, such as validity, relevance and rigour, are also difficult to define. Because concepts are difficult to define, with no agreed definitions in software engineering, we opt for simplicity and use definitions from the Oxford Dictionary [25]. Our terminology is explained below and the relations between concepts are illustrated in Figure 1.

We consider that both producers and consumers of evidence should be concerned with the validity of information, the rigour of the process and the relevance of information and process. The degree of attention that producers and consumers direct to validity, rigour and relevance will likely vary. A

producer of evidence is likely to direct more attention at validity and rigour, whilst a consumer of evidence is likely to direct more attention at the relevance of evidence to their particular purpose.

4.1. Definitions and a reference model

Researchers make observations in research studies, transform those into results, and then interpret those results. The word observation is used here in the general sense for something that has been noticed in any research study. Observations, results and interpretations are all *information*.

We define **evidence** as "The available body of facts or *information* indicating whether a belief or proposition is true or valid." ([25]; emphasis added here). Information takes the status of evidence when a producer or consumer chooses to use that information to support (or negate) a belief or proposition. The producer or consumer has the responsibility to check that it is reasonable to use the information to support (or negate) a proposition.

Credibility is treated as an aspect of quality (we use quality here to mean degree of excellence), while **credible** refers to something being possible to trust, believe or rely upon. We use the phrase **credible evidence** as a way to capture the quality of evidence. Information therefore attains the status of credible evidence when we are able to trust that evidence. For an evidence–based discipline such as software engineering, credible evidence ought to be one of the main research outputs, with an appropriate balance made between the two main characteristics of such output, validity and relevance. The balance between validity and relevance needs to be addressed in each individual case depending on the objective of the research. Credible evidence is therefore an overarching quality aspect of the output of research and it is built up from a combination of validity and relevance, as illustrated in Figure 1 with the solid lines.

Validity is concerned with the information being logically and factually sound. By contrast, **relevance** is concerned with the information being useful for a target group, for example, other researchers or practitioners. Moreover, information may be relevant for a general target group such as practitioners, although not relevant for a specific case. Hence, it may not be fit–for–purpose in a particular situation, although generally relevant for the target group. Because of the significance of **fit–for–purpose**, we explicitly recognise it as a sub–characteristic in Figure 1. Validity also has sub–characteristics, such as reliability. Unlike fit–for–purpose, the sub–characteristics of validity are not significant for our discussion. We therefore

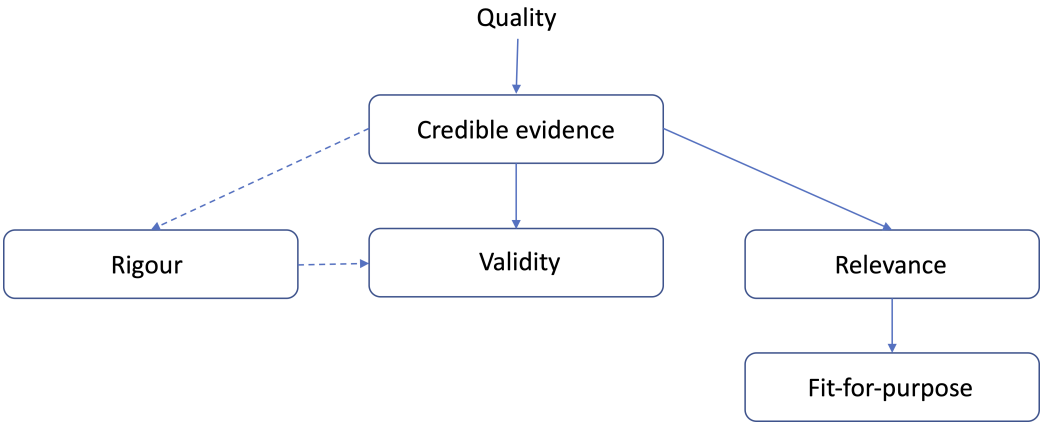

Figure 1: Terminology for different characteristics of quality of evidence.

do not report these sub–characteristics in Figure 1 so as to remain focused on the overarching quality characteristic.

Rigour is concerned with thoroughness and carefulness of the process, i.e., that research is conducted in a rigorous way, e.g., following available research guidelines. Thus, rigour may be used as a direct indicator of validity, and hence as an indirect indicator of credible evidence.

In Figure 1, we use a dotted line to indicate that rigour has a different kind of relationship amongst the concepts. Rigour includes both the appropriate choice of research process and the appropriate application of that process. Thus, rigour supports validity in terms of the conduct of research. Rigour also has sub–characteristics, such as replicability. Again we do not present those sub–characteristics in the model in Figure 1 as they are not significant to our discussion.

We recognise that the terminology in Figure 1 will most likely be interpreted differently when viewed from the different perspectives presented in Table 1, and also that different individuals may have different opinions even when taking the same perspective. Thus, for example, something that is credible evidence for one person may not be credible for someone else; it depends on different perspectives, for example, those of the researcher in contrast to those of a practitioner. Furthermore, different individuals may have different opinions on what constitutes credible evidence given their context, knowledge and experience. The same reasoning holds for the other characteristics in Figure 1.

4.2. Further comments on the definitions

4.2.1. Accumulation of evidence

An individual primary study typically produces a 'piece' of evidence. 'Pieces' of evidence from different primary studies may each be credible but will likely vary in their rigour, validity and relevance. In particular, a primary study may have no direct relevance to software engineering practice but may still be relevant in the sense that the evidence is relevant for research. (A primary study that has *no* relevance to research or to practice would likely be at best classified as irrelevant research.)

As pieces of evidence are properly aggregated, for example with secondary studies, so the research community can begin to accumulate credible evidence (or more strictly, increasingly credible evidence). The proper aggregation of evidence would ideally lead to a body of evidence that has higher validity and is more relevant, i.e., more credible evidence that is also useful to software practice. To move from individual primary studies generating 'pieces' of credible evidence to a body of knowledge is one of the overall objectives of evidence–based software engineering research.

4.2.2. Responsibilities of producers and consumers

We argue that both the producer and the consumer have the responsibility to check that evidence is credible evidence. Our belief is a prescriptive belief rather than a normative belief. Ideally, a producer of information should argue for why that information would constitute credible evidence, e.g., the information has higher validity (because it has been collected and transformed in a sufficiently rigorous way, and it is interpreted as being possible to trust) and it is sufficiently relevant (because that information increases or reduces our belief in some aspect of the world of interest to us). In reality, however, the processes for collecting, analysing and interpreting such information may be faulty, and the arguments themselves may be incomplete or faulty. Similarly, the consumer should also ideally (but may not) argue for why the information is credible evidence. In reality, the consumer may often simply accept that information as credible evidence, because the consumer accepts that information to have been rigorously produced by the producer and because the consumer considers the information to be sufficiently relevant to their interest. In reality, a particular person may be both a producer and a consumer, e.g., a researcher both consumes information produced by researchers in previous studies whilst also producing new information as part of her or his new research study.

4.3. Relating our definitions to previous research

Like Prechelt and Petre's paper [10] (see Section 2.1) we recognise the importance of credibility, however we focus on credible evidence and we relate the characteristics of validity and relevance in a way different to the relationships described in the Prechelt and Petre paper.

With our definitions, credible evidence has two main characteristics, validity and relevance. We treat rigour as an indicator of validity, and hence indirectly related to credible evidence. Thus, we distinguish between the degree of validity of information and the rigour of the correct process for the research to be conducted. For example, Prechelt and Petre [10] write, "Credibility requires a study to embody not just high validity but also *good reporting* ..." (emphasis in original). We argue that a study, as a process of inquiry, should be evaluated for its rigour, with the information (evidence) used by or produced from the study being evaluated for its validity. And both the process of inquiry and the information used in, or produced from, the inquiry should be properly reported.

Like Ivarsson and Gorschek [11] (see Section 2.2), we recognise relationships between rigour, validity, relevance and fit–for–purpose, although (again) we relate these concepts in a way different to the relationships specified by Ivarsson and Gorschek. For example, we do not subsume validity within rigour, and we recognise the concept of credible evidence. Also, Ivarsson and Gorschek's study is focused particularly on technology evaluations and industry relevance, whereas our work is intended to have a more general application.

Finally, Petersen and Gencel [12] (see Section 2.3 for our brief review) discuss the relationship between philosophical worldviews, research methods, types of validity and threats to validity. We do not explicitly consider worldviews in our definition on credible evidence.

4.4. Summary

In summary, rigour concerns the conduct of the research, while validity and relevance concern the evidence output from the research. Evidence may have higher validity but not be relevant, or evidence may have higher validity and be relevant but not be fit–for–purpose. Credible evidence is evidence with higher validity which is relevant and which may – depending on the context – be fit–for–purpose. Producers and consumers have the responsibility to ensure, and assure, that they are producing and consuming credible evidence, i.e., they are conducting their own research rigorously (e.g.,

by carefully conducting and reporting the research process), are producing credible evidence, are aggregating others' research carefully, and are applying evidence in a relevant way.

5. Credible evidence and challenges

In this section, we discuss our primary case and our two secondary cases of published research to illustrate challenges with producing, consuming and disseminating credible evidence. We briefly complement our three cases with five other cases to demonstrate that the primary and secondary cases are not unique in the literature. Our research approach has been summarised in Section 3.

For each case, we map the case to concepts introduced elsewhere in this paper. First, using Table 2, the cases are mapped to the perspectives of Researcher–as–Producer (RaP), Researcher–as–Consumer (RaC) and Practitioner–as–Consumer (PaC), as introduced in Section 1.1. Then, also using Table 2, the cases are mapped to two characteristics of credible evidence, i.e. validity and relevance, as illustrated in Figure 1. Finally, using Table 3, each case and its challenges are mapped to the recommendations we present in the different subsections of Section 6.

Table 2: Descriptive summary of cases.

#	§	**Cases**	**RaP**	**RaC**	**PaC**	**Validity**	**Relevance**
1	5.1	The misinterpretation of data by producers, and its subsequent dissemination	X	x		X	
2	5.2.1	The misinterpretation of evidence by consumers	x	X		X	x
3	5.2.2	Issues in presenting evidence across multiple publications	X	x		x	x
4	5.3.1	Further producer risks for misinterpretations	X	x		X	
5	5.3.2	Further consumer risks for misinterpretations	x	X		X	x
6	5.3.3	Vested interest	X		x	X	
7	5.3.4	Concerns regarding fit–for–purpose of evidence: context and ageing			X		X
8	5.3.5	Claims make it into evidence	x		X	X	

Notes:
"X" denotes primary perspective discussed in the case.
"x" denotes supplementary perspective discussed in the case.

Table 3: Areas of challenges, cases and challenges.

Areas of challenges	§	**Cases**	§	**Challenges**
Researchers as producers	5.1	The misinterpretation of data by producers, and its subsequent dissemination	6.2.1	Conducting studies
	5.3.1	Further producer risks for misinterpretations		
	5.2.2	Issues in presenting evidence across multiple publications	6.2.2	Managing a series of studies
	5.3.3	Vested interest	6.2.3	Addressing vested interest
Researchers as consumers	5.2.1	The misinterpretation of evidence by consumers	6.3.1	Citing evidence
	5.3.2	Further consumer risks for misinterpretations		
Practitioners as consumers	5.3.4	Concerns regarding fit–for–purpose of evidence: context and ageing	6.4.1	Using evidence in practice
	5.3.5	Claims make it into evidence	6.4.2	Avoiding making claims into evidence

5.1. Primary case – Misinterpretation of data and dissemination

5.1.1. Introduction to the case

As mentioned in Section 3, our first case concerns the misinterpretation of a table of data that relates module size and module defect density. The paper reporting and misinterpreting this table of data is the Basili and Perricone [24] paper published in 1984. For convenience, the table of data is reproduced here in Table 4. The paper that first explicitly identified and discussed the *misinterpretation* in the Basili and Perricone paper was the paper by Rosenberg [26], published 13 years later in 1997. There is a paper by Shepperd [27], published in 1988, that expresses concerns about the relationship, however Shepperd's paper does not explicitly identify the problem subsequently identified by Rosenberg. We consider Shepperd's paper later in our discussion.

Table 4: Defects / 1000 Executable lines (from [24])

Module size	**Defects / 1000 lines**
50	16
100	12.6
150	12.4
200	7.6
> 200	6.4

.

The issue that interests us here, and which was identified and described by Rosenberg [26], concerns Basili and Perricone's [24] conclusion that smaller modules have higher defect densities than larger modules. The 'root' problem identified by Rosenberg is that this relationship is due to the mathematical coupling between module size and defect density. This is because both module size and module defect density are derived from lines of code: module size is calculated in terms of the number of lines of code and module defect density is calculated in terms of the number of defects divided by the number lines of code. Thus, as long as the number of lines of code increases faster than the number of defects increase, the relationship will continue to be of the form $y = 1/x$. Hence, the relationship identified by Basili and Perricone [24] is actually explained by mathematical coupling, i.e. the collinearity between the variables. The advice provided by Rosenberg [26] is that "... a compound metric should never be related to any other metric with which it shares a component". Basili and Perricone's misinterpretation of data led to their invalid conclusion, and therefore the invalid evidence that smaller modules have higher defect densities than larger modules. This misinterpretation was presumably also not identified during the peer–review process prior to the paper's publication. In fact, the authors explicitly acknowledge this surprising result in their paper and the editor alludes to this surprising result in a brief editorial comment that accompanies the paper.

The problem of collinearity exists in other types of software engineering studies too. An illustrative example is when productivity is compared with the size of the input. This is problematic since productivity itself is defined as the size of the output divided by the size of the input. Thus, productivity is a compound metric, and it should not be related to size of the input. This problem is pointed out in several papers, for example, by Jørgensen and Kitchenham [21], and in the book by Bossavit [28].

5.1.2. Citation analysis

As briefly discussed earlier, to examine the dissemination of the invalid conclusion in the Basili and Perricone paper [24], and therefore the impact of invalid evidence, the first author conducted a citation analysis of both the Basili and Perricone paper and the Rosenberg paper. The details of that analysis are presented in Appendix A. Here, in the main paper, we focus on the results of that analysis.

Table 5, reproduced from the table in Appendix A for ease of reference, presents the results of the analysis. As explained in Appendix A, 126 pub-

Table 5: Number of citations to the papers by Basili and Perricone, and by Rosenberg.

	Basili & Perricone [24]		**Rosenberg [26]**
Citations	**1984–1997**	**1998–2016**	**1998–2016**
Total	8058 (21)	12029 (50)	3984 (55)
Positive	832 (4)	4119 (11)	45 (2)
Negative	0 (0)	74 (1)	3095 (21)
Vague	197 (1)	1273 (3)	0 (0)
Note: The number of papers generating citations are shown within parenthesis.			

lications citing the Basili and Perricone paper and/or the Rosenberg paper were selected and classified. The citing publications were classified according to four categories (the four rows in Table 5): the total number of citations, the number of citing publications being 'positive' to the relationship (e.g., the paper argued that it corroborated in some way the Basili and Perricone conclusion), the number of citing publications being 'negative' to the relationship (e.g., the paper argued that the Basili and Perricone conclusion was invalid), and the number of citing publications being 'ambivalent' to the relationship (e.g., the paper expresses concern about the Basili and Perricone conclusion but does not reject the relationship outright).

For each cell in Table 5, the cell includes both the number of citations and the number of citing publications responsible for these citations (shown in parenthesis after the number of citations). For example, the 21 papers investigated in the first interval are themselves cited in total 8058 times.

As the table indicates, this analysis was performed for two time intervals: the period between the publication of Basili and Perricone's paper in 1984 and the publication of Rosenberg's paper in 1997, and the period between the year after the publication of Rosenberg's paper and the year prior to when our citation analysis was conducted. Both time intervals are considered for the classification so as to ensure the fair treatment of citing publications to the two papers. The sum of citing publications and citations in rows 2–4 is not equal to the first row, because the two source papers can be cited for reasons other than the relationship between module size and module defect density. Our concern here is only with publications that cite the relationship observed by Basili and Perricone.

5.1.3. Interpretation of the citation analysis

Table 5 shows that Basili and Perricone's paper, and in particular their evidence about the relationship between size and defect density, motivated considerable further research, and shows that researchers continued to accept the invalid conclusion (as credible evidence) even after Rosenberg's [26] paper. Researchers presumed, and searched for, an optimal module size, referred to as the Goldilocks Principle by Compton and Whitrow [29]. Two particularly relevant cases are, first, Hatton's [23] use of the relationship to support claims about software reliability; and second, Ostrand *et al.*'s series of papers [30, 31, 32, 33, 34, 35, 36] synthesising evidence across multiple papers and studies. We explicitly consider these two cases in the next subsection of our paper. Both cases highlight additional challenges to the production and consumption of credible evidence.

Returning to Table 5, the table shows that no publication is critical of the evidence in Basili and Perricone's paper before the paper by Rosenberg. As noted earlier, there is one publication that appears to be unsatisfied with the Basili and Perricone evidence. The 'unsatisfied publication' is a paper by Shepperd, published in 1988 [27], and indicated by the entry of '(1)' in the 'vague' category of Table 5, for the time interval 1984–1997. Shepperd writes, "...there is disagreement as to whether error density is an appropriate means of size normalisation since module size and error density do not appear to be independent." Note that Shepperd's paper was published four years after the Basili and Perricone paper and almost 10 years *before* the Rosenberg paper. Shepperd's paper provides for a counter–factual, what–if scenario: *what if* Shepperd's paper had been able to explicitly raise the concerns that Rosenberg' paper subsequently raised — might that have 'saved' the community 10 years of misinterpreted research effort, the proliferation of invalid evidence, and a multitude of misplaced citations? Our what–if scenario is intended to emphasise an important function of evidence–based research: to challenge accepted evidence.

Table 5 also shows that, when comparing citations to the two source papers, the number of citations to the Basili and Perricone paper are substantially higher. The average number of citations to publications citing the Basili and Perricone paper in the second interval (1998–2016) is more than three times as high as those citing the Rosenberg paper (cf. 12029/50 and 3*3984/55).

Moreover, eleven highly cited papers write positively about the evidence

reported by Basili and Perricone *after* the paper by Rosenberg has been published. This may be a consequence of the abundance of publications, which makes it hard to find the 'right' paper. But, in any case, the propagation of the problematic evidence continues. Furthermore, two papers — one by Stensrud *et al.* [37], the other by Koru and El Emam [38] — although not highly cited, discuss Rosenberg's observation in such a way that the discussion may be interpreted as disagreeing with Rosenberg. Stensrud *et al.* [37] question the generalisability of Rosenberg's observation in the context of Magnitude of Relative Error (often denoted MRE), although they quote Rosenberg's remark that his observation requires that defects do not grow more than linearly with the number of lines of code, which at face value sounds reasonable. Koru and El Emam [38] refer to the paper by Basili and Perricone [24] and their speculation that the findings may be a consequence related to interface defects. The speculation is inline with the finding presented by Koru and El Emam [38], i.e. that smaller modules have proportionally higher coupling. They state that the speculation by Basili and Perricone is plausible given their observation concerning coupling. As discussed below, Koru and El Emam [39] have also authored a paper highlighting the concerns raised by Rosenburg, although the paper is published one year earlier than the paper discussed here.

The encouraging news is that a number of papers highlight the concerns raised by Rosenberg and these papers are themselves reasonably well–cited.

Finally, it is worth noting that only four of the investigated publications cite *both* source papers, i.e., [40, 41, 42, 39]. The ways in which these four publications address the contradiction between the Basili and Perricone conclusions and the Rosenberg arguments are quite different. Fenton and Neil [40] present the evidence in Basili and Perricone's paper as an observation, and the discussion about Rosenberg is primarily related to collinearity. The contradiction between the two papers is noted, but not really highlighted. El Emam *et al.* [41] report on the evidence by Basili and Perricone, point to correlation being difficult, and refer to other papers. They are not very explicit about the potential issue. Andersson and Runeson [42] state that Rosenberg questions the validity of comparing size and defect density, but the contradiction between the papers is not highlighted. Finally, Koru *et al.* [39] are very clear about the mathematical coupling pointed out by Rosenberg explicitly recognising this in the abstract of their paper and demonstrating it in the body of the paper.

5.1.4. Summary

To summarise the above:

1. Researchers unintentionally misinterpreted their data, derived an invalid conclusion from that misinterpretation, and subsequently published invalid evidence.
2. The researchers' misinterpretation was presumably not identified during the peer–review process.
3. The invalid conclusion (though not realised at the time, of course) motivated considerable subsequent research for many years.
4. Even when the conclusion was demonstrated to be invalid, researchers continued to cite the evidence.

Setting aside the mathematical coupling, the outcome that smaller modules have lower defect density is hard to explain from a software development perspective. Software design principles prescribe that software engineers should divide software into smaller logical components and to not write one large piece of software. The conclusions from the Basili and Perricone paper [24] seem to contradict software design principles, e.g., the divide–and–conquer strategy. Thus, when finding results that are perceived to be counter–intuitive, it is essential to challenge those results and to try to understand the reasons behind the findings. Many studies have shown that size is strongly correlated with most measures used to predict defects, and hence it would be counter–intuitive if it was actually true that smaller components have proportionally more defects. When observing results that are counter–intuitive, we should be very cautious. The reasoning above illustrates how results may be interpreted wrongly, and hence that there is the need for extra care with counter–intuitive results.

5.2. Secondary cases – Consistent presentation of evidence

5.2.1. The misinterpretation of evidence by consumers

In the previous section, we discussed the misinterpretation of data by evidence–producers, leading to invalid evidence. Whilst as researchers we produce our own evidence, we also cite the research and evidence of others, and in so doing we can unintentionally misinterpret others' evidence. When analysing the citations to the paper by Basili and Perricone [24], we identified a separate issue concerning the interpretation of Basili and Perricone's evidence.

In the paper by Basili and Perricone [24], the authors study defect density. They do not make any statements about reliability. It is therefore unfortunate when the evidence in the Basili and Perricone paper (despite its other problems) is then misapplied when being cited. An example of this misapplication can be found in a paper by Hatton [23], who writes: "Their most important conclusion was, once again, that small components were proportionately more unreliable than the larger components ...". The relationship between module size and module defect density has now been (mis)interpreted as evidence for a relationship between module size and reliability. Furthermore, given the wide dissemination of the original evidence, there is the risk that the new meaning could also be widely promulgated, i.e. a relationship between module size and reliability. In our experience, the risk comes from our observation that citations have a tendency to generate further citations, since each citation helps make a paper more visible to the community.

Overall, the case illustrates the challenge to the consumer of evidence – and also the *responsibility* on the consumer of evidence – to interpret other's evidence correctly.

5.2.2. Presenting evidence consistently across multiple publications

Authors sometimes produce a series of papers that are closely related to each other, e.g., a series of papers that analyse defect data, but use different analyses techniques, or different releases of the software, or different subsets of releases of a system. Such series of papers may be highly valuable, e.g., because they can accumulate evidence. But it is crucial that the reader, as a consumer of that evidence, is able to understand the relationships between the papers in the series, so that they can make proper use of the evidence produced from the studies reported in the papers. Where the series of papers, and their respective studies and evidence, are not properly explained, the consumer may not be able to properly interpret the evidence. In the worst case, it may not even be possible to determine that the same system has been studied. This limits the value of the evidence presented in the papers.

When analysing the Basili and Perricone [24], and the Rosenberg [26] papers in Section 5.1, we identified a series of papers by Ostrand, Weyuker and Bell [30, 31, 32, 33, 34, 35, 36] (Bell is not a co–author of [30]). This series of seven papers is interesting because it illustrates the challenges of presenting evidence across a series of papers.

Six of the Ostrand *et al.* papers refer, in general terms only, to the

Basili and Perricone [24] paper, whilst only one paper [30] – the first paper, published in 2002 –has a more specific reference to the Basili and Perricone [24] paper. That one paper [30] reports that its findings support the results of Basili and Perricone concerning the relationship between module size and defect density.

The seven papers were all published after Rosenberg's paper, during the period 2002 – 2008, yet none of the papers cite Rosenberg's paper and all seven papers cite the Basili and Perricone paper. Our focus for this case is not on the lack of citation to Rosenberg's [26] paper, or on the visibility of the Basili and Perricone paper [24] through the citations (though the latter may indirectly contribute to making the conclusions in the Basili and Perricone paper more visible). The focus here is on the way in which new evidence produced across the seven papers is presented.

Table 6 presents a summary of the systems studied in each paper. Table 7 presents a summary of the relationships between the seven papers. The evidence in each paper is clearly presented. However, it is difficult to synthesise the evidence across the seven papers, which means that the value of having a series of closely–related papers is not exploited to the extent possible (and even more useful for the readers).

Table 6: Summary of the seven papers by Ostrand, Weyuker and Bell.

Paper	Ref	Date	S1R	S2R	S3M	S4R
1	[30]	2002	13			
2	[31]	2004	17	9		
3	[32]	2005	17	9		
4	[33]	2006			18M	
5	[34]	2007	17	9	24M	
6	[35]	2007				35
7	[36]	2008	17	9	24M	35

A brief description of the systems (as interpreted by the current authors:
S1: Inventory system.
S2: Provisioning system.
S3: Voice response system.
S4: Maintenance support system.
In the column headings, R refers to releases and M refers to months.
Paper 7 presents Systems S1, S2, S3 and S4 as described in papers 5 and 6.

The seven papers are clearly related to each other and the authors seem to mostly refer to the papers as one would expect. Not all papers have the same research questions which, again, is as one would expect: different papers would typically be motivated by different research questions.

But given the overlap in systems and releases, and given the differences

Table 7: Relationships of seven papers by Ostrand, Weyuker and Bell.

Paper	Ref	Citations and relationships
1	[30]	is the first paper in the series, presenting information on 13 releases of a system, S1.
2	[31]	refers to Paper #1 and explicitly states that it is the same system (S1). Paper #2 refers to 12 releases in Paper #1, but Paper #1 includes 13 releases.
3	[32]	seems to be a journal extension of Paper #2, including S1 and S2, but does not cite Paper #2. Paper #3 again (incorrectly) refers to 12 releases in Paper #1.
4	[33]	presents the study of a new system, S3, and adds developer information. The paper refers to previous work, primarily Paper #2 and Paper #3, but also includes a general reference to Paper #1.
5	[34]	is focused on comparison of results between systems. The paper includes the two systems, S1 & S2, from Papers #1–3. Furthermore, it includes the system reported in Paper #4 with additional development, i.e., 24 months instead of 18 months. Paper #5 includes references to Papers #1–4.
6	[35]	presents a study of a new system, S4, and reports a high level comparison with Systems S1–S3. Paper #6 includes references to Papers #1–4.
7	[36]	presents and investigates developer information in detail for S4. In particular, the paper adds the number of developers to the models built in the authors' previous research, and performs an analysis for S4. Paper #7 refers to Papers #1, #3 and #4. Paper #2 is listed in the reference list of Paper #7, but it is not referenced in the paper.

in research questions, it becomes quite difficult to determine new evidence, to distinguish new from existing evidence and, in general, to distil the evidence. There is (or appears to be) nothing formally wrong with the papers, however as one example it would be helpful if a paper was clear whether it was an extension of a previous paper, so that the reader as a consumer of evidence can better understand the status of the evidence. The importance of clarifying the relationship between papers is apparent for [31] and [32] (see papers 2 and 3 in Table 7).

When analysing the references between the papers in more detail, we found the following. References are mostly used to motivate the current paper and to highlight differences between the investigations. Furthermore, many references relate to findings, systems, and methodology in the other papers, for example, "... preliminary findings can be found in reference N".

Of most relevance to the current discussion are references to previous observations. Unfortunately few examples are present in the set of seven papers. References often relate to specific systems but insufficient details are provided about the systems to promote a general understanding of the reasons for different or similar findings. Instead, results are sometimes contrasted at a general level. Most significantly, the results from the different papers are not synthesised, i.e. there is no building of new knowledge as the

series of papers progresses. Thus, the issue here is primarily a presentation concern, and at the end an unrealised opportunity for the authors to present even more valuable results to the consumers of the research.

Overall, this case illustrates the challenge of carefully distinguishing evidence, and then of the need to synthesise evidence, as one conducts more studies in a series, for example, studies that concern similar research questions for different systems and releases. Phrased another way, each paper provides a piece or pieces of evidence, but these pieces are not then aggregated into a body of more credible evidence.

The challenge of synthesising evidence across a series of papers is closely related to the challenge of the reappearance of data. In many situations, the same data is the basis for more than one paper, and hence it is essential that it is clear to the reader as the consumer of evidence that it is the same data; or where there are differences in the data, these differences are described and explained.

The challenge of the reappearance of data has become more visible with the adoption of systematic literature reviews. For example, reviewers can find it difficult to know whether the data, and related evidence, reported in a paper is *new* data, or whether the data and its analysis is *reused*, or whether the data is reused but with new analysis. This makes it difficult to know whether and how to include a paper in a review or whether to exclude the paper.

5.3. Complementary cases concerning credible evidence

5.3.1. Further producer risks for misinterpretations

To complement the cases in Sections 5.1-5.2, there is also the need to be cautious when some important information is unknown, since such unknown information may also increase the risk of misinterpreting results. Thus, the producer of evidence should carefully report important aspects related to the evidence put forward. For example, the CHAOS report [43] has been quoted to motivate research, while others have been more critical towards the report, for example, [44, 45]. The main concern raised is the lack of information about definition of measures and the methodology used.

Furthermore, to avoid misinterpretations, it is important that the most appropriate statistical methods or other analysis methods are applied to ensure that the evidence generated is correctly interpreted and presented. Several authors have contributed with papers outlining the use of statistical

methods, for example, Kitchenham *et al.* [15] and others have provided recommendations for analysis of qualitative data [46] and synthesis of findings [47, 48].

5.3.2. Further consumer risks for misinterpretations

We complement the case in Section 5.2.1 with a case from Dybå *et al.* [49]. Dybå *et al.* [49] criticise papers on contextual factors, for example, a paper by Clarke and O'Connor [50]. The paper by Clarke and O'Connor [50] presents a reference framework (also called a checklist). The paper presents a number of factors in the reference framework. The paper does not state that the reference framework (or checklist) should be used in its entirety. Dybå *et al.* [49] observe that the number of combinations of the factors in the reference framework would be extremely large. Thus, Dybå *et al.* present a potentially misleading impression of the paper. The reference framework or checklist is meant to ensure that important contextual factors are not forgotten. Thus, it is up to the user of the reference framework to decide which contextual factors are important in each specific case.

Overall, the case further illustrates the challenge to the consumer of evidence – and also the *responsibility* on the consumer of evidence – to interpret other's results correctly and to communicate a correct impression of related work.

5.3.3. Vested interest

Evidence may be challenged when a producer or a consumer of the evidence has a vested interest in the outcome. Even if there is no intentional bias, there may still be unconscious bias, and this threatens a fair comparison. For vested interest, we briefly discuss one case from previous research, i.e., Shepperd *et al.*'s [51] systematic review of cost estimation.

The issues with researcher bias, and in particular its relation to expertise, is highlighted by Shepperd *et al.* [51] in their systematic review of cost estimation. Shepperd *et al.* write: "Surprisingly we find that the choice of classifier has little impact upon performance (1.3 percent) and in contrast the major (31 percent) explanatory factor is the researcher group. It matters more who does the work than what is done." Shepperd *et al.*'s [51] case illustrates the challenges concerning credible evidence, and the difficulty to make fair comparisons when conducting evaluations.

5.3.4. Fit–for–purpose of evidence: context and ageing

An important consideration when producing evidence is the fit–for–purpose of that evidence. Evidence may have higher validity, and may also be relevant (e.g., the evidence relates to a topic of interest to the consumer) but may not be practical because it does not apply, or cannot be applied, to the specific software development context.

Software cost estimation is a challenging research area in software engineering. Software cost estimation has occurred since the inception of software engineering as a discipline. Boehm conducted some of the pioneering work with COCOMO [52] and the literature on the subject is extensive. Kitchenham *et al.* [53] identified seven systematic literature review concerning software cost estimation. However, is the evidence on the topic useful? In many cases the evidence is not useful since the estimates do not necessarily reflect the development. The latter may be due to changes in requirements or by removing some features before delivery. Thus, it becomes questionable if the original estimate can be compared with the outcome. Furthermore, some studies indicate that both estimate–increase and estimate–decrease exist [54]. In some of these studied cases, the changes may be explained by intentional distortions, for example, due to management pressure. Given that we, at least in some cases, cannot trust that the estimate and the outcome may be compared, how do we know which data is useful?

One specific aspect of the fit–for–purpose of evidence, and the data it is built upon, is the ageing of evidence and data. Ageing relates to the interval in time between when the data was collected, and the evidence derived, and when the evidence from that data is being applied. One challenge is to understand when data is getting too old. The field of software development moves very quickly, and hence some data is not useful any longer, for example, data connected to specific technologies, programming languages and tools that are no longer used. Thus, it becomes essential to understand the lifespan of evidence in different areas of software engineering.

5.3.5. Claims make it into evidence

In the book *The Leprechauns of Software Engineering* [28], Bossavit presents several cases of anecdotal claims that have become accepted as credible evidence in software engineering and adopted as such in practice. (A *leprechaun* is part of Irish folklore, and hence a leprechaun in the context of software engineering concerns folklore and myths. Thus, it relates to myths where the evidence is, at best, questionable.) We briefly consider one of the

cases highlighted by Bossavit, i.e., Boehm's cone of uncertainty.

According to Bossavit, the cone of uncertainty was originally an *opinion* of Barry Boehm's, an opinion that Boehm *illustrated* with a diagram. That opinion was subsequently presented, in a paper by another author, as "empirically validated". Thus, according to Bossavit [28] (see p. 17), the original opinion has been reformulated to become credible evidence. This "evidence" may then be applied by practitioners, i.e., as a consumer of evidence. The described case is very well aligned with the problem in Section 5.1 where Basili and Perricone present an observation, which then is spread and also reformulated as being about module reliability as discussed in Section 5.2.1.

Bossavit provides further examples in [28], where he also states a pledge to researchers ([28], p. 138) concerning writing research papers and hence moving towards factual claims, which is fully aligned with evidence–based research in software engineering.

5.4. Summary

In summary, Section 5 has presented one primary case, two secondary cases, and five complementary cases, pointing to issues related to credible evidence. We have classified these cases in relation to producers and consumers, and to validity and relevance (see Table 2). Furthermore, we have used the cases to highlight challenges concerning credible evidence (see Table 3), and considered some of the consequences of propagating invalid or questionable evidence.

6. A framework and recommendations for credible evidence

In Section 1 we introduced perspectives on evidence. In Section 2, we presented related work, focusing in particular on rigour, validity, relevance and credible evidence in a software engineering context. In Section 4 we defined terminology. In Section 5, we presented cases to illustrate challenges. In the current section, we bring together these preceding sections into a framework, and complement that framework with a set of recommendations concerning challenges to producing, consuming and disseminating credible evidence.

6.1. A framework for credible evidence

The framework is presented in Figure 2. Credible evidence is at the centre, and the perspectives of producer and consumer are shown toward the

top and the bottom of the figure respectively. Consistent with Section 1.1, the main producer is assumed to be a researcher while both researchers and practitioners are consumers of research. To the left, we have the scientific approach and to the right a focus on the relevance to context and its relation to practice.

Credible evidence brings together a researcher's interest in science, e.g., validity, with a practitioner's interest in usefulness, e.g., relevance and fit–for–purpose (though fit–for–purpose is not shown in Figure 2). Rigour may be viewed as an indicator of validity (illustrated with a dotted line in Figure 2), since it is expected that research conducted in a rigorous way is more likely to produce results with higher validity.

In an ideal situation, it should be possible to conduct research in a rigorous way to obtain results that have higher validity and are relevant for practice (and fit–for–purpose). Unfortunately, this ideal is too rarely attained, and hence it is often necessary to decide on the main intended consumers of the research, i.e., to choose between intended consumption by researchers or intended consumption by practitioners. As a consequence, it becomes essential to focus on the characteristics of credible evidence that are most important for that type of consumer. The focus relates to the choice of research methods, which is further discussed in, for example, Wohlin and Aurum [55], and which is beyond the scope of the framework shown in Figure 2.

In summary, the framework combines the two main characteristics of credible evidence with the two perspectives of producer and consumer. The framework is intended to structure the concept of credible evidence and its relations to research and practice. The overall goal is that the framework should work as a tool, for both producers and consumers of research, to move towards more credible evidence in software engineering.

Drawing on the cases discussed in Section 5 and the framework introduced above, we present in the following sections a set of recommendations for dealing with challenges to the production, consumption and dissemination of credible evidence. The recommendations are intended to help producers and consumers of evidence to better position the credibility of that evidence. Our recommendations are intended to complement existing checklists and guidelines and so help to ensure that we both produce credible evidence and are able to assess whether or not the evidence presented is credible.

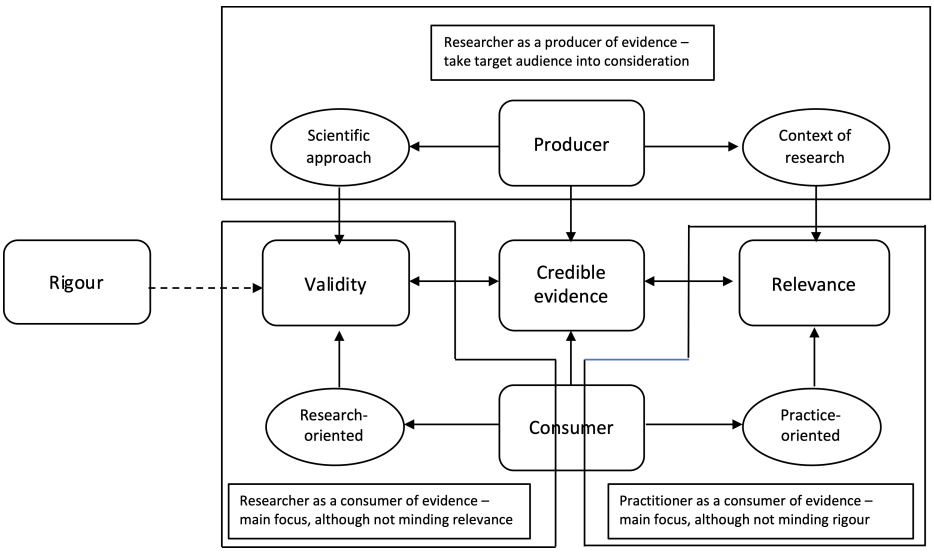

Figure 2: A framework for credible evidence.

6.2. Recommendations – researchers as producers of evidence

6.2.1. Conducting studies

The analysis of the case in Section 5.1, concerning the relationship between module size and defect density, is intended to emphasise the need for the research community to be very careful when publishing and disseminating evidence.

Furthermore, the discussion in Section 5.3.1 concerning the CHAOS 1994 report highlights the need to present definitions of measures and research methods in sufficient detail that the consumer can trust the rigour of the process and therefore the validity of the evidence. As mentioned above, the rigour of the research is often an indicator of the validity of the evidence. We propose the following recommendations to help researchers produce more credible evidence:

1. *Design* research studies to produce credible evidence. This includes having a strong research team, potentially including different expertise depending on the study. Furthermore, the choice of research method needs to be carefully considered [55], as should the number of data points that ought to be obtained for credible evidence. Finally, the context of the study and, if appropriate, the selection of suitable subjects for the study both need to be considered.

2. *Select* appropriate evidence because not all available "evidence" can or should be taken at face value. Selecting appropriate evidence is a delicate task since evidence should be selected based on the above framework, and not based on fitness towards a specific objective to "prove" something. In the process of selecting evidence, the person making the selection has the obligation to argue for her or his choice of evidence *and* for the inferences made from that choice of evidence. One very important role of the peer–review process is to hold the person selecting evidence accountable for their choices and arguments. This includes both their arguments for evidence and their arguments with evidence.
3. *Interpret* the evidence carefully. In particular, counter–intuitive evidence should be scrutinised very carefully. The case in Section 5.1 illustrates that problematic evidence may spread, and hence it is important to be careful when publishing evidence and when citing a paper. If problems are observed, we should be open about the concerns and express the concerns in a constructive way. It may also be advisable to contact the authors of papers to get more details and open a dialogue between researchers.
4. *Report* methods used for the research, so as to ensure that the consumer of the evidence understands how the evidence was obtained. The appropriateness and rigour of the research methods used are often viewed as an indicator of the validity of the evidence.
5. *Ensure* true relationships are reported. We need to be careful to avoid confounding factors, and hence report cause and effect relationships. Particular care should be taken with correlation.
6. *Use* the most appropriate analytical methods for the data, and then the evidence, generated by a research study. This is far from easy and researchers are not always in agreement on the most appropriate methods, and recommendations change as lessons are learned. For example, in Kitchenham *et al.* [15], Kitchenham explains how she originally advised the use of box plots but now recommends the use of kernel density plots to visualise data.

6.2.2. Managing a series of studies

As discussed in Section 5.2.2, a series of papers puts special challenges on the authors to ensure that the full potential of the evidence from the

series can be properly exploited. To move in the direction of usefulness of the evidence, we offer the following advice:

1. *Describe* carefully the context of the phenomenon (or phenomena) under study. This is to ensure that practitioners can better evaluate whether the evidence is fit–for–purpose, and researchers can better evaluate whether the evidence is relevant to their research.
2. *Contrast* clearly *all* already–published papers in a series so that consumers can better evaluate the evidential contribution of each paper e.g., when a new paper is an extension of a previous paper.
3. *Relate* the evidence from different papers to each other to ensure that the full value of the evidence generated from the series of studies and papers is properly communicated.

In Section 5.2.2, Tables 6 and 7 provide indicative examples of how to contrast papers and evidence.

6.2.3. Addressing vested interest

It may be impossible to avoid a vested interest when comparing different alternatives, as discussed in Section 5.3.3. The vested interest may include both being the inventor of a new method, technique or tool [56], or having disproportionate amounts of expertise in *some*, but not all, of the alternatives being evaluated *cf.* [51].

Based on these challenges, we propose the following recommendations to help researchers produce credible evidence:

1. *Build* a diverse research team for research studies to compensate for vested interest, for example, when conducting an evaluation, include researchers who have not been involved in the development of the new method, technique or tool that is being evaluated.
2. *Ensure* that the team includes expertise for the different methods, techniques and tools to be compared in an evaluation, even when not including an inventor of something new. Thus, even when comparing existing methods, techniques and tool, it is essential to include complementary expertise.
3. *Design* comparative studies in a fair way. For example, to have a well–described reading technique and compare it with a simplistic checklist is not a fair comparison.

4. *Report* carefully what you have done to enable public scrutiny, e.g., for replication. Where there is a potential vested interest, it is even more important to describe all details to make the evidence transparent and hence more credible.

6.3. Recommendations – researchers as consumers of evidence

6.3.1. Citing evidence

Interpretation of published evidence by researchers (and indeed by practitioners) needs to be handled with care. It is essential that evidence is not distorted (cf. Section 5.2.1) or that citations are made, to specific statements or parts, out of its context (cf. Section 5.3.2). This leads to the following recommendations:

1. *Avoid* confirmation bias by carefully examining papers before citing them. If a paper is published (having been reviewed by other researchers), and in particular if a paper is written by well–established researchers with an excellent reputation, it is crucial to not just accept findings, as illustrated above. Thus, if replicating a study, it is essential to start from the beginning and do the groundwork, and then potentially observe that the previous findings ought to be challenged.
2. *Ensure* correctness by carefully evaluating the available evidence before citing a paper. If evidence is questionable, it is important that the evidence is not cited without an appropriate discussion of its credibility. If there is doubt about the evidence consider contacting the authors.
3. *Cite without distortion*, i.e. cite carefully and ensure that the papers are interpreted correctly, for example, statements cannot be made about software reliability based on defect density. If in doubt, it is good to contact the authors of the original paper. Furthermore, to intentionally or unintentionally cite a paper incorrectly is a disservice to the research community, and as we have shown can have an considerable long–term impact on the direction of research.
4. *Cite without sub–optimal selection*, i.e., citations should be based on the full information provided in a paper. Thus, we should be careful when we cite an isolated statement, or cite a part of a paper out of its context. It may lead the reader of a paper to get the wrong impression of the paper being cited. Citations to a paper do not, unfortunately, distinguish between the validity of different pieces of evidence in that paper, for example, as in the Basili and Perricone paper [24], which

contains both the problematic relationship between module size and defect density, as well as other results.

6.4. Recommendations – practitioners as consumers of evidence

6.4.1. Using evidence in practice

Fit–for–purpose is a challenge, since it is highly dependent on the context. Furthermore, the cost estimation discussion, in Section 5.3.4, highlights that the estimate may not be fully comparable with the outcome for different reasons as discussed in Section 5.3.4.

We propose the following recommendations for practitioners who will use the research outputs:

1. *Describe* the key contextual factors that are considered most relevant to ensuring the evidence is fit–for–purpose. We need to better understand the influence of the context in relation to when evidence is fit–for–purpose.
2. *Report* the context carefully when publishing papers. We need to become better in describing the context of our studies so that others can make a judgement whether or not the evidence is relevant for them as consumers, independently of them being researchers or practitioners.
3. *Ensure* correct comparisons of data, e.g., we need to know that an estimate is truly comparable with an outcome. If not, the comparison does not make sense and hence the evidence generated does not become credible.

6.4.2. Avoiding making claims into evidence

In Section 5.3.5, we briefly discussed Bossavit's [28] book and highlighted a case of how a claim has been taken as evidence. The book also includes a pledge to authors in software engineering to provide the best possible basis for their factual claims, i.e. with credible evidence. The pledge implies several recommendations. Some of the more important ones are summarised here:

1. *Cite* only papers you have read, and ensure traceability to what is cited.
2. *Check* carefully that the cited papers supports the claims made, i.e. do not overstate evidence or allow invalid evidence to spread by citations.
3. *Look* for all types of evidence to avoid confirmatory bias.

7. Conclusions

We conclude our paper with a consideration of further directions for research in this area, and a brief review of the objectives of our paper.

7.1. Further research

There are several directions in which our work could be extended.

In Section 2 we recognised, but chose to set aside for this paper, that practitioners can be producers of evidence. Illustrative examples of the practitioner–as–producer are Fred Brooks and his book, *The Mythical Man–Month*, and Tom DeMarco and Timothy Lister's book, *PeopleWare*. One area for further research is therefore to investigate the challenges in the production of credible evidence by practitioners, and the consumption and subsequent dissemination of that evidence by practitioners and researchers. A distinctive challenge is that practitioners rarely use a recognised process of inquiry; indeed if they did use such a process then, according to our categorisation in Table 1, the practitioner would actually be conducting research. Instead of assessing the rigour of an inquiry process, the 'production' of practitioner–generated evidence may need to be assessed in terms of the expertise and impartiality of the practitioner–as–producer. As we noted earlier, the rigour to the process of an inquiry is only an indicator of credible evidence. We can therefore distinguish the validity and relevance of practitioners' *evidence* from the expertise and impartiality of the practitioner as the *producer* of that evidence.

A second direction for research, one that is connected with our model in Figure 1, is the connection between credible evidence and fit–for–purpose. As Prechelt and Petre write, "Different purposes require different *standards of evidence*." ([10], p. 20; emphasis in original). This direction of research could be extended to consider the sub–characteristics of the constructs in Figure 1.

The main focus of our paper has been on the (mis)interpretation of evidence and the subsequent publication and dissemination of such evidence. A third direction for research — and an area that has already been explored considerably — is therefore the importance of a rigorous methodology, i.e., applying a methodology rigorously helps us to assure that we are not misinterpreting evidence, but nevertheless misinterpretations can still take place.

A final direction for research is to evaluate and extend our recommendations so as to help ensure wider coverage of the challenges confronting

the production, consumption and dissemination of credible evidence by researchers and practitioners.

7.2. Review of objectives

In Section 1, we stated the following objectives:

1. To identify some of the challenges confronted by producers and consumers when producing, consuming and disseminating evidence.
2. To identify some of the consequences that arise from the production, consumption and dissemination of invalid or otherwise questionable evidence.
3. To provide guidance to the community on how to address some of these challenges.

For the first objective, we presented and discussed cases from prior research. Our cases illustrated a range of challenges that need to be carefully handled:

- Conducting studies
- Managing a series of studies
- Addressing vested interest
- Citing evidence
- Using evidence in practice
- Avoiding making claims into evidence

For the second objective, we used our cases to demonstrate some of the consequences that arise from disseminating invalid or misinterpreted evidence. The most prominent case was discussed in Section 5.1 where Table 5 shows that Basili and Perricone's [24] paper, and in particular their evidence about the relationship between size and defect density, motivated considerable further research, and that researchers continued to accept the invalid conclusion as credible evidence even after Rosenberg's [26] paper was published. Researchers presumed, and searched for, an optimal module size, referred to as the Goldilocks Principle [29]. We also suggested a counter–factual what–if scenario, using a paper by Shepperd [27], to speculate on the

potential 'savings' of research effort if the invalid evidence had been identified and acted upon sooner.

For our third objective, we provide a framework and a series of recommendations in Section 6 to advise on the challenges listed above.

In addressing our objectives, the paper contributes a terminology for reasoning about credible evidence, a set of cases, a set of challenges, and a framework and recommendations for addressing the challenges.

Acknowledgements

We would like to express our gratitude to the reviewers and the editors for valuable comments and input that helped us improve the paper.

References

[1] B. Kitchenham, D. R. Jeffery, C. Connaughton, Misleading metrics and unsound analyses, IEEE Software 24 (2007) 73–78.

[2] B. Kitchenham, L. Madeyski, P. Brereton, Problems with statistical practice in human-centric software engineering experiments, in: Proceedings of the 23rd Conference on Evaluation and Assessment on Software Engineering (EASE), ACM, 2019, pp. 134–143.

[3] R. P. Reyes, O. Dieste, E. R. Fonseca, N. Juristo, Statistical errors in software engineering experiments: A preliminary literature review, in: Proceedings IEEE/ACM 40th International Conference on Software Engineering (ICSE), IEEE, 2018, pp. 1195–1206.

[4] J. Soldani, D. A. Tamburri, W.-J. Van Den Heuvel, The pains and gains of microservices: A systematic grey literature review, Journal of Systems and Software 146 (2018) 215–232.

[5] J. S. Molléri, K. Petersen, E. Mendes, Survey guidelines in software engineering: An annotated review, in: Proceedings of the 10th International Symposium on Empirical Software Engineering and Measurement (ISESE), ACM, 2016, p. 58.

[6] S. E. Hove, B. Anda, Experiences from conducting semi-structured interviews in empirical software engineering research, in: Proceedings of the 11th International Software Metrics Symposium (Metrics), IEEE, 2005, p. 23.

[7] J. Kontio, L. Lehtola, J. Bragge, Using the focus group method in software engineering: Obtaining practitioner and user experiences, in: Proceedings of the 3rd International Symposium on Empirical Software Engineering (ISESE), IEEE, 2004, pp. 271–280.

[8] C. Treude, M. P. Robillard, Augmenting API documentation with insights from Stack Overflow, in: Proceedings IEEE/ACM 38th International Conference on Software Engineering (ICSE), IEEE, 2016, pp. 392–403.

[9] D. A. Schum, The Evidential Foundations of Probabilistic Reasoning, Northwestern University Press, 2001.

[10] L. Prechelt, M. Petre, Credibility, or why should I insist on being convinced?, in: A. Oram, G. Wilson (Eds.), Making Software: What Really Works, and Why We Believe It, O'Reilly Media, Inc., Sebastapol, CA, 2010, pp. 17–34.

[11] M. Ivarsson, T. Gorschek, A method for evaluating rigor and industrial relevance of technology evaluations, Empirical Software Engineering - An International Journal 16 (2011) 365–395.

[12] K. Petersen, C. Gencel, Worldviews, research methods, and their relationship to validity in empirical software engineering research, in: Proceedings of the Joint Conference of the 23rd International Workshop on Software Measurement and the 8th International Conference on Software Process and Product Measurement, IEEE, 2013, pp. 81–89.

[13] D. Lo, N. Nagappan, T. Zimmermann, How practitioners perceive the relevance of software engineering research, in: Proceedings of the 10th Joint Meeting on Foundations of Software Engineering, ACM, 2015, pp. 415–425.

[14] H. Zhang, M. A. Babar, P. Tell, Identifying relevant studies in software engineering, Journal of Information and Software Technology 53 (2011) 625–637.

[15] B. Kitchenham, L. Madeyski, D. Budgen, J. Keung, P. Brereton, S. Charters, S. Gibbs, A. Pohthong, Robust statistical methods for empirical software engineering, Empirical Software Engineering - An International Journal 22 (2017) 579–630.

[16] G. A. Liebchen, M. Shepperd, Data sets and data quality in software engineering, in: Proceedings 4th International Workshop on Predictor Models in Software Engineering, ACM, 2008, pp. 39–44.

[17] G. Liebchen, M. Shepperd, Data sets and data quality in software engineering: Eight years on, in: Proceedings of the 12th International Conference on Predictive Models and Data Analytics in Software Engineering, ACM, 2016, p. 7.

[18] R. K. Yin, Case Study Research and Applications – Design and Methods, SAGE, 2017.

[19] P. Runeson, M. Höst, Guidelines for conducting and reporting case study research in software engineering, Empirical software engineering 14 (2009) 131.

[20] S. MacDonell, M. Shepperd, B. Kitchenham, E. Mendes, How reliable are systematic reviews in empirical software engineering?, IEEE Transactions on Software Engineering 36 (2010) 676–687.

[21] M. Jørgensen, B. Kitchenham, Interpretation problems related to the use of regression models to decide on economy of scale in software development, Journal of Systems and Software 85 (2012) 2494–2503.

[22] C. Wohlin, Guidelines for snowballing in systematic literature studies and a replication in software engineering, in: Proceedings of the 18th Conference on Evaluation and Assessment in Software Engineering, ACM, p. 38.

[23] L. Hatton, Reexamining the fault density component size connection, IEEE Software 14 (1997) 89–97.

[24] V. R. Basili, B. T. Perricone, Software errors and complexity: An empirical investigation, Commununications of the ACM 27 (1984) 42–52.

[25] Oxford University Press, Oxford dictionary, https://www.lexico.com/en, 2020. Accessed: 2020-06-30.

[26] J. Rosenberg, Some misconceptions about lines of code, in: Proceedings of the 4th International Software Metrics Symposium (Metrics), IEEE, 1997, pp. 137–142.

[27] M. Shepperd, A critique of cyclomatic complexity as a software metric, Software Engineering Journal 3 (1988) 30–36.

[28] L. Bossavit, The Leprechauns of Software Engineering - How Folklore Turns into Fact and What to Do about It, Leanpub, 2015.

[29] B. Compton, C. Withrow, Prediction and control of Ada software defects, Journal of Systems and Software 12 (1990) 199–207.

[30] T. J. Ostrand, E. J. Weyuker, The distribution of faults in a large industrial software system, in: Proceedings of the 12th International Symposium on Software Testing and Analysis (ISSTA), ACM, 2002, pp. 55–64.

[31] T. J. Ostrand, E. J. Weyuker, R. M. Bell, Where the bugs are, in: Proceedings of the 14th International Symposium on Software Testing and Analysis (ISSTA), ACM, 2004, pp. 86–96.

[32] T. J. Ostrand, E. J. Weyuker, R. M. Bell, Predicting the location and number of faults in large software systems, IEEE Transactions on Software Engineering 31 (2005) 340–355.

[33] R. M. Bell, T. J. Ostrand, E. J. Weyuker, Looking for bugs in all the right places, in: Proceedings of the 16th International Symposium on Software Testing and Analysis (ISSTA), ACM, 2006, pp. 61–72.

[34] E. J. Weyuker, T. J. Ostrand, R. M. Bell, Using developer information as a factor for fault prediction, in: Proceedings of the 3rd International Workshop on Predictor Models in Software Engineering (PROMISE), IEEE, 2007, p. 8.

[35] T. J. Ostrand, E. J. Weyuker, R. M. Bell, Automating algorithms for the identification of fault-prone files, in: Proceedings of the 17th International Symposium on Software Testing and Analysis (ISSTA), ACM, 2007, pp. 219–227.

[36] E. J. Weyuker, T. J. Ostrand, R. M. Bell, Do too many cooks spoil the broth? Using the number of developers to enhance defect prediction models, Empirical Software Engineering - An International Journal 13 (2008) 539–559.

[37] E. Stensrud, T. Foss, B. Kitchenham, I. Myrtveit, A further empirical investigation of the relationship between MRE and project size, Empirical Software Engineering - An International Journal 8 (2003) 139–161.

[38] A. G. Koru, K. El Emam, The theory of relative dependency: Higher coupling concentration in smaller modules, IEEE Software 27 (2010) 81–89.

[39] A. G. Koru, D. Zhang, K. El Emam, H. Liu, An investigation into the functional form of the size-defect relationship for software modules, IEEE Transactions on Software Engineering 35 (2009) 293–304.

[40] N. E. Fenton, M. Neil, A critique of software defect prediction models, IEEE Transactions on Software Engineering 25 (1999) 675–689.

[41] K. El Emam, S. Benlarbi, N. Goel, W. Melo, H. Lounis, S. N. Rai, The optimal class size for object-oriented software, IEEE Transactions on Software Engineering 28 (2002) 494–509.

[42] C. Andersson, P. Runeson, A replicated quantitative analysis of fault distributions in complex software systems, IEEE Transactions on Software Engineering 33 (2007) 273–286.

[43] The Standish Group, The CHAOS Report (1994), `https://www.standishgroup.com/sample_research_files/chaos_report_1994.pdf`, 1994. Accessed: June 30, 2020.

[44] M. Jørgensen, K. Moløkken-Østvold, How large are software cost overruns? A review of the 1994 CHAOS Report, Journal of Information and Software Technology 48 (2006) 297–301.

[45] J. L. Eveleens, C. Verhoef, The rise and fall of the CHAOS Report figures, IEEE Software (2009) 30–36.

[46] C. B. Seaman, Qualitative methods in empirical studies of software engineering, IEEE Transactions on Software Engineering 25 (1999) 557–572.

[47] D. S. Cruzes, T. Dybå, Recommended steps for thematic synthesis in software engineering, in: Proceedings of the 5th International Symposium on Empirical Software Engineering and Measurement (ESEM), IEEE, 2011, pp. 275–284.

[48] X. Huang, H. Zhang, X. Zhou, M. A. Babar, S. Yang, Synthesizing qualitative research in software engineering, in: Proceedings of the 40th International Conference on Software Engineering (ICSE 18), ACM, 2018, pp. 1207–1218.

[49] T. Dybå, D. I. Sjøberg, D. S. Cruzes, What works for whom, where, when, and why?, in: Proceedings of the 6th International Symposium on Empirical Software Engineering and Measurement (ESEM), ACM, 2012, pp. 19–28.

[50] P. Clarke, R. O'Connor, The situational factors that affect the software development process: Towards a comprehensive reference framework, Journal of Information Software and Technology 54 (2012) 433–447.

[51] M. Shepperd, D. Bowes, T. Hall, Researcher bias: The use of machine learning in software defect prediction, IEEE Transactions on Software Engineering 40 (2014) 603–616.

[52] B. Boehm, Software Engineering Economics, Prentice Hall, Lebanon, Indiana, USA, 1981.

[53] B. Kitchenham, O. P. Brereton, D. Budgen, M. Turner, J. Bailey, S. Linkman, Systematic literature reviews in software engineering: A systematic literature review, Journal of Information and Software Technology 51 (2009) 7–15.

[54] A. Magazinius, S. Börjesson, R. Feldt, Investigating intentional distortions in software cost estimation - an exploratory study, Journal of Systems and Software 85 (2012) 1770–1781.

[55] C. Wohlin, A. Aurum, Towards a decision-making structure for selecting a research design in empirical software engineering, Empirical Software Engineering - An International Journal 20 (2015) 1427–1455.

[56] V. R. Basili, S. Green, O. Laitenberger, F. Lanubile, F. Shull, S. Sørumgård, M. V. Zelkowitz, The empirical investigation of

perspective-based reading, Empirical Software Engineering - An International Journal 1 (1996) 133–164.