Information and Software Technology, Volume 95, March 2018, pp. 62-74
DOI: 10.1016/j.infsof.2017.10.017

Reporting Systematic Reviews: Some Lessons from a Tertiary Study

David Budgen[a,*], Pearl Brereton[b], Sarah Drummond[a], Nikki Williams[c,1]

[a]*Durham University, School of Engineering & Computing Sciences, Durham DH1 4LA*
[b]*Keele University, School of Computing & Maths, Staffordshire ST5 5BG*
[c]*Cranfield University, Centre for Electronic Warfare, Information & Cyber, Defence Academy of the United Kingdom, Shrivenham SN6 8LA*

ACCEPTED MANUSCRIPT

Abstract

Context: Many of the systematic reviews published in software engineering are related to research or methodological issues and hence are unlikely to be of direct benefit to practitioners or teachers. Those that are relevant to practice and teaching need to be presented in a form that makes their findings usable with minimum interpretation.

Objective: We have examined a sample of the many systematic reviews that have been published over a period of six years, in order to assess how well these are reported and identify useful lessons about how this might be done.

Method: We undertook a tertiary study, performing a systematic review of systematic reviews. Our study found 178 systematic reviews published in a set of major software engineering journals over the period 2010-2015. Of these, 37 provided recommendations or conclusions of relevance to education and/or practice and we used the DARE criteria as well as other attributes related to the systematic review process to analyse how well they were reported.

Results: We have derived a set of 12 'lessons' that could help authors with reporting the outcomes of a systematic review in software engineering. We also provide an associated checklist for use by journal and conference referees.

Conclusions: There are several areas where better reporting is needed, including quality assessment, synthesis, and the procedures followed by the reviewers. Researchers, practitioners, teachers and journal referees would all benefit from better reporting of systematic reviews, both for clarity and also for establishing the provenance of any findings.

Keywords:
Systematic review, reporting quality, provenance of findings

1. Introduction

The idea of adapting the use of secondary studies (systematic reviews) to form a tool of empirical software engineering was first proposed in 2004 [1]. Since then, they have become a well established tool for empirical research.

However, what may easily be overlooked is that the motivation for using a systematic review in software engineering usually differs from those that occur in other disciplines, such as health, education and the social sciences. For those disciplines, both systematic reviews and the primary studies that form their inputs are commonly sponsored and commissioned by government and research agencies to support practice and policy-making [2]. This influences both the topics that are studied as well as the way that the outcomes are reported.

In software engineering the funding for such studies (when available) is more likely to be from research grants and the choice of topic is apt to be driven by the interests of the researchers involved. Hence systematic reviews in software engineering are more likely to be concerned with identifying research practices, often taking the form of mapping studies [3, 4]. Many also appear to be undertaken to underpin study for a PhD [5], with the focus of the research questions being upon research trends or research practice.

In 2011 we undertook a tertiary study (a systematic review of systematic reviews) to identify how well the systematic reviews then available could be used as a source of material to help inform introductory teaching about software engineering (and hence by implication, could provide useful knowledge to underpin software engineering practice) [6]. For convenience we will refer

*Corresponding Author
Email addresses: david.budgen@durham.ac.uk (David Budgen), o.p.brereton@keele.ac.uk (Pearl Brereton), sarah.drummond@durham.ac.uk (Sarah Drummond), nikki.williams@cranfield.ac.uk (Nikki Williams)

[1]The work reported in this paper was undertaken when Nikki Williams was employed by Keele University.

Published by Elsevier. This is the Author Accepted Manuscript issued with: Creative Commons Attribution Non-Commercial No Derivatives License (CC:BY:NC:ND 4.0). The final published version (version of record) is available online at DOI10.1016/j.infsof.2017.10.017:. Please refer to any applicable publisher terms of use.

to this as ETS1 (Education Tertiary Study 1) in this paper. More recently, we have been undertaking the task of updating this tertiary study to cover systematic reviews published up to the end of 2015 (we will refer to this as ETS2). In doing so, we have taken the opportunity to refine and extend our analysis of the quality of the processes reported for these reviews and the provenance for their findings.

For ETS2 we have extracted more detailed data than we did in ETS1, and this has required that we examine the papers reporting the systematic reviews in greater detail, both in terms of the nature of the 'body of evidence' found, and also of how the outcomes were reported. Extracting this body of evidence has revealed that the reporting of secondary studies is often incomplete, and not always well organised, as well as providing some examples of good reporting practices.

It is the way that a systematic review is *reported* that forms the topic for this paper, with the pedagogical implications arising from ETS2 being reported separately in [7]. For this paper we have taken a subset of the systematic reviews being used in ETS2 (those published in the period 2010-2015), and undertaken some further data extraction and analysis in order to address the following (supplementary) research question.

> "*For systematic reviews that address topics relevant to software engineering education and practice, how well are the procedures and outcomes of the review reported, and what lessons about good reporting practice can be derived from them?*"

We refer to the resulting analysis as STS1 (Supplementary Tertiary Study 1) and it is the findings from STS1 that we report in this paper.

We have also made use of the lessons derived from our analysis in the way that we report the findings of STS1. Although a tertiary study of this nature is essentially a mapping study, most of the reporting issues are common to all systematic reviews, whatever the form.

Why do we consider it important to undertake this study? Firstly, despite the dominance of research-focused reviews, there are some that address issues of potential value for the wider software engineering community. However to use their findings appropriately it is necessary to be informed about their *provenance*, and about any limitations that might apply to the findings. To do this requires that the process followed in the review as well as the outcomes should be clearly reported.

Secondly, researchers are not always reporting potentially useful findings in a manner that makes them readily accessible to the wider community. Improving the way that outcomes are described and reported can help to influence the future of software engineering as a discipline, by aiding with making the processes of both teaching and practice more *evidence-informed*. And obviously this applies to systematic reviews that address research issues too.

This paper provides a distinct and novel contribution with regard to how to *report* the findings of a systematic review. To do so it employs the same model as we used in an earlier (and widely-cited) paper [8], where the experiences of *conducting* systematic reviews were codified by structuring them as a set of 'lessons for practice'. We hope that the lessons from this study can likewise help guide both researchers (in reporting their work) and referees (in assessing its suitability for publication).

The rest of this paper is structured as follows. In the next section we address background material relevant to four key aspects: the way that systematic reviews are performed; how their quality can be assessed; reporting practices for systematic reviews; and how they can support what is taught. Section 3 describes our research method and the design of our tertiary study. Section 4 reports upon the conduct of the study, and in Section 5 we present our findings and seek to derive lessons from these, as well as considering the limitations of our study. Finally we provide some recommendations for improving reporting practices and a checklist for reviewers.

2. Background

The use of systematic reviews in software engineering is now well established and well documented. Guidelines on how to perform a systematic review in software engineering were originally formulated by Kitchenham in 2004, and updated in 2007 [9]. A further update has also been provided in a book addressing the use of systematic reviews in software engineering [10]. Guidelines for performing mapping studies have also been formulated by other researchers [3].

2.1. Forms of Systematic Review

A systematic review aims to locate all studies that contain material of relevance to its research question, and to synthesise the outcomes of those studies considered relevant. For that reason they are sometimes termed *secondary studies*, while the studies forming their input are termed *primary studies*. A research-oriented variation that has been widely used in software engineering is the *mapping study*, which is a form of systematic review that does not seek to perform any significant element of synthesis, but instead categorises the

ACCEPTED MANUSCRIPT

primary studies against some framework or model, in order to identify what research has been undertaken, and possibly, where there are gaps in this [3, 11].

Tertiary studies, which seek to synthesise or 'map' secondary studies, also take a variety of forms. A *broad* tertiary study is more like a mapping study and seeks to identify and categorise systematic reviews, possibly around some thematic issue. Three broad tertiary studies summarised all of the systematic reviews published up to the end of 2009 [12, 13, 14]. The rapidly increasing number of systematic reviews then made performing such studies to be both a very large undertaking and probably also one of diminishing value, and so later tertiary studies have tended to be more constrained in scope. Our own studies on the usefulness of systematic reviews for teaching (ETS1 and ETS2) can be considered to be broad tertiary reviews conducted around the theme of meeting educational needs. More focused tertiary studies usually look at systematic reviews related to a software engineering topic (such as testing) or a research practice, and seek to categorise or synthesise their findings. A good example is the tertiary study looking at research synthesis [15].

Our paper reports the findings from a focused tertiary study (STS1).

2.2. Quality Assessment of Systematic Reviews

The degree of confidence that we can place upon the findings from a systematic review (their *provenance*) will depend upon factors such as how thoroughly it was performed; the quality of the outcomes from the primary studies included in it; and the domain knowledge of the researchers performing the review. So being able to assess the quality of a systematic review in an organised manner is an important function for a tertiary study.

In the field of clinical medicine, where such confidence is a particularly important aspect, and where many of the ideas about systematic reviews have been pioneered, a widely used assessment scheme is that known as DARE[2] (Database of Attributes of Reviews of Effects). The original DARE assessment was based upon four questions, later extended to five, which in their most abstract form are as follows.

1. Are the review's inclusion and exclusion criteria described and appropriate?
2. Is the literature search likely to have covered all relevant studies?
3. Did the reviewers assess the quality/ validity of the included studies?
4. Were basic data/ studies adequately described?
5. Were the included studies synthesised?

Other quality assessment schemes do exist and have been used for software engineering studies, but for ETS1 and ETS2 we chose to adopt the DARE scheme (this was also used in the broad tertiary studies). However, we should note that this is concerned with the systematic review process—and that assessing the quality of the primary studies should be a function of the review itself. DARE is only concerned with *whether* such an assessment has been done, not *how well* it has been done, and there are no agreed criteria for assessing the abilities of a review team.

When employing DARE, a commonly-used convention is one of scoring each question on a three-point scale: yes (1), partly (0.5), no (0). Hence the maximum possible score is 5.0. Scoring is undertaken by two researchers, who base their independent scores on a more detailed interpretation of the above criteria, applying these criteria to the procedures reported in the systematic review. After comparing their scores for the individual questions they then resolve any differences to produce an 'agreed score'.

2.3. Reporting Systematic Reviews

In reporting the findings from a systematic review, there are two major aspects that need to be addressed. The first of these is to fully report the process that was followed for the review—which should include the activities identified in the DARE criteria above, as well as the details about how any synthesis was performed. The thoroughness with which this is done can both aid later updating of a systematic review as well as provide readers with confidence about the findings.

The second is how to present the actual findings. This aspect is less well-defined in the sense that the procedures for determining how the outcomes should be interpreted are not well established [16]. Even for clinical studies, the way that the activities forming what is usually termed *Knowledge Translation* (KT) should be organised is still an issue of debate. The main goal of KT is to provide *guidelines for practice* derived from the findings of a review, but of course, the interpretation involved needs to encompass many other factors, including the organisational context within which the guidelines will be used.

In the context of clinical practice, Khan et al. highlight the following needs that should be addressed by any recommendations provided by a review [17].

[2] http://www.crd.york.ac.uk/CRDWeb/AboutPage.asp

ACCEPTED MANUSCRIPT

- Recommendations should convey a clear message and following them in practice should be as simple as possible.

- Potential users need to be informed about how credible (trustworthy) the recommendations are, where their credibility stems partly from the strength of evidence provided from the review, as well as other factors, including the thoroughness of the review process.

We discuss the issues of reporting and provenance more fully in Section 3.

2.4. Using Systematic Reviews in Education

The findings from ETS2 that are relevant to teaching about software engineering are addressed more fully elsewhere [7]. Hence this subsection is confined to identifying what we regard as constituting a systematic review considered to be useful for teaching, and hence appropriate for inclusion in our study.

In [6] we argued that systematic reviews could be used to make teaching about software engineering more *evidence-informed* by providing information for both teachers and students about what works, how well it might work, and in what context it is likely to work best. We also suggested that the outline of knowledge about software engineering topics provided in the *ACM/IEEE Curriculum Guidelines for Undergraduate Degree Programs in Software Engineering* provided a useful scoping and categorisation scheme through the summary of core topics defined in the SEEK (Software Engineering Education Knowledge). The original curriculum guidelines were published in 2004, and updated in 2014[3].

The process undertaken for revising the curriculum guidelines involved widespread consultation with both educators and practitioners [18]. So although our main interest has been in the use of these findings for supporting education, it is reasonable to assume that they will also be relevant to practitioners.

To select papers for ETS2 (and hence STS1) we have therefore used the current curriculum guidelines to help identify where there is material in a systematic review that could be used to support introductory teaching of software engineering topics to undergraduates (and that might also benefit students on taught masters programmes as well as practitioners).

[3] http://securriculum.org

3. Research Method

In this section we discuss the procedures adopted for this study, as encapsulated in the *review protocol* that we prepared beforehand, and provide some rationale for their choice. We also explain our interpretation of the quality assessment criteria used.

3.1. Scope of the tertiary study

As explained in the introduction, this analysis was performed partly by using material collected for a study seeking to identify systematic reviews containing material that could be used to support and inform teaching of introductory software engineering (ETS2). ETS2 extends an earlier tertiary study on material for teaching (ETS1) [6], and includes an additional analysis of the *provenance* of the findings from each systematic review.

Figure 1 shows the basic relationships between the three tertiary studies we performed, with emphasis upon the role of each one. In this figure we have not attempted to include details of the data collection in order to avoid an excess of detail, but should note here that STS1 makes use of some of the data extracted for ETS2 along with some further, more detailed, data extraction that addressed issues concerned with reporting.

All of the authors of this paper have extensive experience of teaching software engineering topics at different levels, and as indicated, we used the SEEK from the 2014 Curriculum Guidelines as a general guide to suitability. The inclusion/exclusion criteria used are summarised in Table 1.

All decisions about inclusion/exclusion were based on analysis by two of the reviewers, working in different pairings to help minimise bias. For a paper to be useful for teaching, we considered it necessary for there to be a clear link between the data and any conclusions or recommendations provided by the original authors, and we discuss this issue more fully in Section 5.

3.2. Searching for systematic reviews

For this study on reporting (STS1) we used a subset of the systematic reviews found in the updated study on teaching material (ETS2), confining our analysis to the systematic reviews published in the period January 2010 to December 2015. This was on the basis that by 2010 the procedures for systematic reviews in software engineering could be considered as well established and widely available. Researchers had also had time to become familiar with performing systematic reviews.

To find systematic reviews, we conducted a search through the five major software engineering journals listed in Table 2. These were the ones also used in our

ACCEPTED MANUSCRIPT

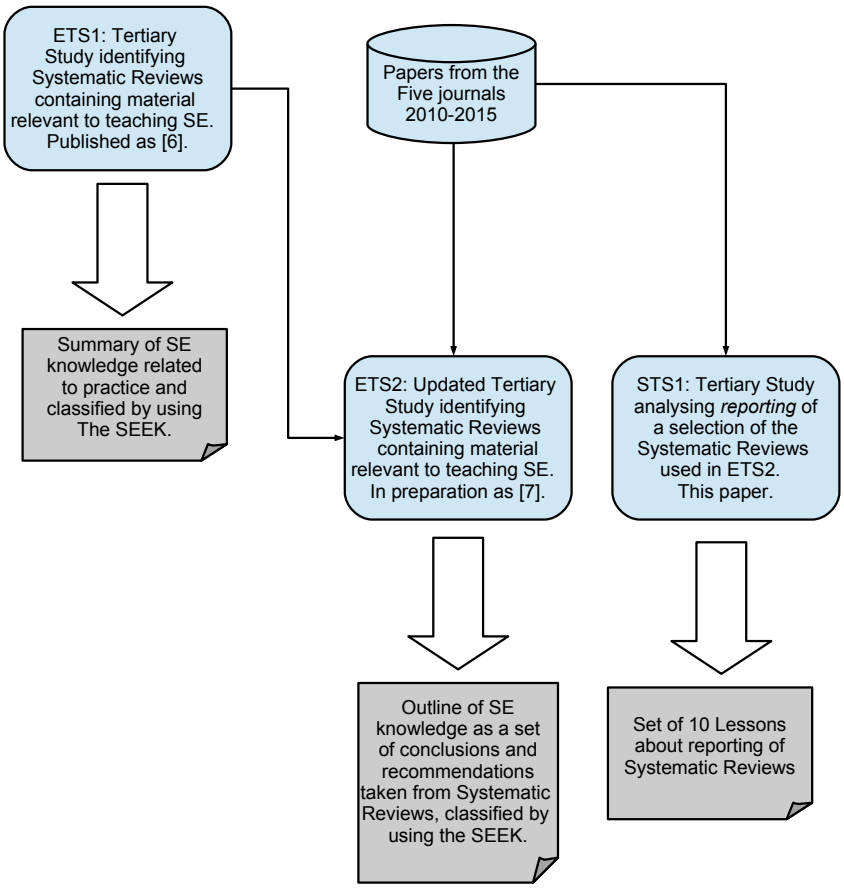

Figure 1: The relationship between the tertiary studies

previous study [6] and included one journal (*Information & Software Technology*) that had established a regular section for such reviews.

Our rationale for using a restricted search was that these journals were considered to be the major publishers of systematic reviews in software engineering and hence able to provide a representative set of systematic reviews. Experience from our earlier study also indicated that conference publications were often mapping studies, and if not, a study of any significance was likely to be extended for journal publication. We also included a small number of candidate studies from other journals, suggested to us by other researchers.

Selection involved a two-stage process. In the first stage, two members from the team, working in different pairings, examined each review to determine whether it met (or appeared to meet) the inclusion/exclusion criteria. In the second stage, two members of the team independently undertook a detailed quality assessment and data extraction as described below. During the first stage, we only excluded a paper if both reviewers considered it should be excluded.

3.3. Quality Assessment

For quality assessment of the systematic reviews selected as meeting the inclusion/exclusion criteria, we used the interpretation of the DARE criteria shown in Table 3. Each value of 'yes' was scored as 1.0, a 'partly' as 0.5 and a 'no' as 0.0. Again, scoring was undertaken by two members of the team, and any differences in the scores were resolved by discussion in order to reach an agreed value.

3.4. Data extraction

After some pilot exercises, we devised an instrument for data extraction for use with ETS2 that recorded the following aspects of a systematic review, where available.

- Bibliographical information (title, authors, publication details).

ACCEPTED MANUSCRIPT

Table 1: Inclusion/Exclusion Criteria used in the Tertiary Study

Inclusion Criteria	
I1.	The paper is published in an issue of a journal within the period 1 January 2010 and 31 December 2015.
I2.	The topic of the paper is appropriate for introductory teaching and falls within the scope of the SEEK.
I3.	The paper contains conclusions or recommendations relevant to teaching that are explicitly supported by the results.
Exclusion Criteria	
E1.	Systematic reviews addressing research trends.
E2.	Systematic reviews addressing research methodological issues.
E3.	Mapping studies with no synthesis of data.
E4.	Systematic Reviews on topics not deemed relevant to introductory teaching of software engineering.

Table 2: Journals used as sources of Systematic Reviews

Journals Searched (2010-2015)
Empirical Software Engineering (EMSE)
IEEE Transactions on Software Engineering (TSE)
Information & Software Technology (IST)
Journal of Systems & Software (JSS)
Software–Practice & Experience (SPE)

- Our scores for the DARE criteria (as described above).
- Data about any quality assessment performed on the primary studies (number of items in the checklist, whether this was derived from other checklists, the actual questions used).
- The size and nature of the *body of evidence* used in the review (numbers and types of study).
- The *context* for the body of evidence (details of participant types, period covered by search, search engines used, details of any manual searches, use of snowballing, number of studies retained at each stage of inclusion/exclusion).
- Any *conclusions* that are reported or could be derived from the paper, together with information about how these were linked to the data (the body of evidence).
- Any *recommendations* that are reported or could be derived, together with information about how these were linked to the body of evidence.

We also made provision to record details of where this information was to be found, and in what form, and for any other points thought to be relevant.

While the specific conclusions and recommendations from individual systematic reviews are not considered further in this paper (belonging correctly to the more pedagogical analysis of ETS2), our use of these terms does need to be explained here. This is because we report on how many of them were identified for each review, as well as discussing the issues that were encountered in extracting them.

From the pilot studies we concluded that, while few studies presented any explicit recommendations, or even conclusions that were relevant to teaching and practice, these could often be extracted from the paper. We also considered it to be useful to make a distinction between these as follows.

- A *conclusion* is considered to be knowledge about the topic that a teacher or a student might find helpful when gaining an understanding about the topic, but which does not provide explicit advice about good or poor practice.
- A *recommendation* is knowledge that could help with making decisions about practice. Where possible, the degree of confidence in a recommendation should also be associated with some measure of its *strength*, derived from the quality of the relevant elements from the body of knowledge.

Overall, we only included those conclusions and recommendations that were related to the topic of the paper and to practice and omitted those concerned with research (nearly every systematic review concludes that there is a need for more and better primary studies!). For both conclusions and recommendations a condition for inclusion was that the reviewers could identify some explicit link to the paper's body of knowledge to justify their inclusion. Wherever possible, we also contacted the original authors of a paper to check that we had extracted these correctly.

In addition, we subsequently performed a further data extraction to help address the research question for STS1, which consisted of the following items.

- How the quality scores were used in the review.
- The form(s) of *synthesis* used, as identified by the original authors or ourselves, and using the categories for synthesis forms described in [15].

ACCEPTED MANUSCRIPT

Table 3: Interpretation of the DARE Criteria used for the tertiary study

Criterion	Score	Interpretation
Inclusion & exclusion	yes	The criteria used are explicitly defined in the paper.
	partly	The inclusion/exclusion criteria are implicit.
	no	The criteria are not defined and cannot be readily inferred.
Search coverage	yes	The authors have searched four or more digital libraries and included additional search strategies OR identified and referenced all journals addressing the topic of interest.
	partly	Searched three or four digital libraries with no extra search strategies OR searched a defined but restricted set of journals and conference proceedings.
	no	Searched up to two digital libraries or an extremely restricted set of journals.
Assessment of quality	yes	The authors have explicitly defined quality criteria and extracted them from each primary study.
	partly	The research question involved quality issues that are addressed by the study.
	no	No explicit quality assessment of individual papers has been attempted.
Study description	yes	Information is presented about each paper.
	partly	Only summary information is presented about individual papers.
	no	The results for individual studies are not specified.
Synthesis of studies	yes	The authors have performed a meta-analysis or used another form of synthesis for all the data of the study.
	partly	Synthesis has been performed for some of the data from some of the primary studies.
	no	No explicit synthesis has been performed (as in a mapping study).

- The procedures employed by the original authors for performing tasks related to inclusion/exclusion.
- The procedures employed by the original authors for performing quality assessment (where such an assessment was undertaken).

4. Conduct of the Study

We begin by summarising the process followed for our tertiary study and describe the resulting set of systematic reviews. We then look at the values obtained for each of the five DARE criteria, and also describe the characteristics of the conclusions and recommendations extracted from this set of papers.

Figure 2 summarises the overall review process we followed and indicates the number of papers that were included at each stage.

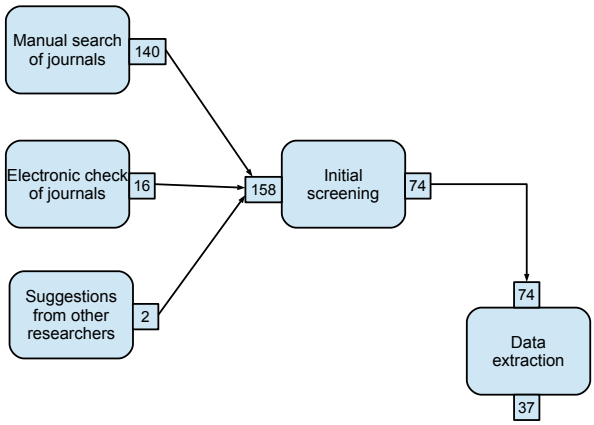

Figure 2: Overview of the selection process

4.1. Study selection

The manual search process was undertaken by one member of the team (DB). It involved reading through the contents pages of the five journals for the period 2010-2015, as determined by the research protocol described in Section 3. While most systematic reviews could be readily identified from their titles or associated meta-data, a few also required inspection of the abstract.

As a check on this, an electronic search was also performed by an independent researcher. The search was done in two stages. Initially the Scopus digital library was used to perform a forward citation analysis of six papers that discussed the principles of EBSE and systematic reviews. This search extended the search reported in [19] and was performed in April 2016. The

ACCEPTED MANUSCRIPT

papers identified as systematic reviews or mapping studies for the five relevant journals were compared with the papers found by the manual search. However, the search identified a large number of false positives. So for papers published in 2015, Scopus was searched using the terms: TITLE-ABS-KEY ("systematic literature review" OR "systematic review" OR "systematic mapping study" OR "mapping study") AND DOCTYPE (ar OR re) AND PUBYEAR = 2015 AND (LIMIT_TO (SUBJAREA, "COMP")). The results from the search were sub-setted to select studies from each of the five journals and the papers that were identified as mapping studies and systematic reviews were compared with the papers found by the manual search. The second search took place in May 2016.

All of the reviews were allocated an index number, beginning from 121 (following on from the total of 120 systematic reviews identified in the three broad tertiary studies, which covered the period to the end of 2009). The systematic reviews found in the five journals added another 156 reviews, which together with two recommended by researchers from other journals (one of which fell outside the period covered in this study) resulted in a total of 278 systematic reviews as candidates for the pedagogical study, and a subset of 158 systematic reviews for the period covered by this study. The rest of this subsection is only concerned with that subset of papers.

The next stage was that of initial selection which was mainly concerned with whether or not a study was a systematic review that addressed a potentially relevant topic. This was performed by all four reviewers, working in different pairings, organised on a random basis, except for those papers in which two of us were authors (PB and DB), with those having to be assessed by the other two reviewers. Initial selection involved a 'quick read' of key sections to determine whether or not the study met the inclusion/exclusion criteria. (We might add that the titles of many of the papers were unreliable, papers titled as systematic reviews were sometimes little more than a mapping study, while occasionally a paper described as a mapping study would involve some degree of synthesis.) This left a total of 74 studies for more detailed analysis and data extraction.

The third and final stage involved a process of data extraction that followed the plan set out in Section 3 and again used random pairing of team members. This led to the exclusion of 37 papers, either on the basis that it was not possible to identify clear links between the data and any conclusions or recommendations provided; or because on closer examination we could not identify specific conclusions or recommendations. This left a final set of 37 systematic reviews considered to provide material that could be used to support teaching and practice, and published within the period 2010-2015. It is this set of systematic reviews that provide the basis for the analysis presented in this paper.

As a check on the reliability of the interpretations made during our data extraction, we contacted the authors of each of the systematic reviews that were included in the final set, and asked them to comment on our interpretation of the outcomes (conclusions and recommendations). We heard back from the authors of 23 papers, none of whom suggested other than minor changes to wording.

Table 4 provides a summary of the 37 systematic reviews. For each one, we have provided the following information.

- The *index number* assigned to this review in our studies. This can be used to assist with cross-reference between our different tables.
- The bibliographic reference for the systematic review.
- The period covered by the systematic review. (Where there was an 'open' start date, we have described this as "to ⟨year⟩".)
- The year of publication.
- The values assigned by our analysis to the five DARE questions, and the total quality score from these.
- The number of *conclusions* and *recommendations* extracted from a review. Care should be used when interpreting these numbers, since some conclusions and recommendations are relatively simple ones, while others are rather more complex, arising chiefly from the way that the data for a particular review has been synthesised.

The data extraction performed for ETS2 (and hence STS1) was more comprehensive than that which we undertook for ETS1 [6]. Also, for the earlier study we used an 'extractor-checker' procedure, whereas for this one all extraction was performed independently by two of the team, who then resolved any differences to produce an agreed dataset for the review. (In [10] there is an explanation of why the latter is now recommended in preference to the former.) As a result, we excluded one of the papers that had been included in the earlier study.

Table 5 provides the totals of papers involved at each stage in this process. These are further broken down by

ACCEPTED MANUSCRIPT

Table 4: Summary details of the systematic reviews included in this study

Index	Ref.	Period covered	year publ.	primary studies	DARE Scores						Concl.	Rec.
					i/e	srch	qual	desc	syn	Total		
121	[20]	2000-2007	2010	59	1.0	0.5	0.0	0.5	1.0	3.0	8	0
123	[21]	unclear	2010	19	1.0	1.0	0.0	1.0	0.5	3.5	5	0
124	[22]	1970-2007	2010	130	1.0	0.5	1.0	0.5	0.5	3.5	1	0
126	[23]	1989-2006	2010	79	1.0	1.0	1.0	1.0	1.0	5.0	1	0
130	[24]	1997-2008	2010	22	1.0	0.5	1.0	1.0	1.0	4.5	5	0
134	[25]	to 3/2005	2011	30	1.0	1.0	1.0	1.0	1.0	5.0	0	5
135	[26]	1980-2008	2011	72	1.0	1.0	1.0	0.0	0.0	3.0	1	8
138	[27]	to 2009	2011	38	1.0	1.0	1.0	1.0	0.5	4.5	4	5
146	[28]	2000-2010	2011	70	0.5	0.0	0.5	1.0	0.5	2.5	3	0
150	[29]	to 6/2010	2011	39	1.0	1.0	1.0	0.0	0.5	3.5	4	0
154	[30]	1995-2009	2012	20	1.0	1.0	0.0	0.0	0.0	2.0	1	0
155	[31]	2000-2010	2012	36	1.0	1.0	1.0	1.0	0.5	4.5	10	0
157	[32]	to 2/2011	2013	27	1.0	0.5	0.5	1.0	1.0	4.0	1	0
160	[33]	to 4/2009	2012	42	1.0	0.5	0.0	0.5	0.5	2.5	3	0
161	[34]	1993-2011	2012	47	1.0	1.0	1.0	1.0	0.5	4.5	0	3
167	[35]	2006-2011	2013	82	1.0	1.0	1.0	0.5	0.5	4.0	3	0
174	[36]	unclear	2014	87	1.0	0.5	1.0	1.0	0.0	3.5	1	1
175	[37]	to mid-2008	2011	98	1.0	0.5	1.0	0.0	1.0	3.5	2	5
193	[38]	to 7/2010	2013	100	1.0	1.0	0.0	1.0	1.0	4.0	2	0
197	[39]	to 10/2011	2013	106	1.0	1.0	1.0	0.5	1.0	4.5	7	3
205	[40]	2000-2011	2014	41	1.0	1.0	1.0	1.0	0.5	4.5	3	0
215	[41]	to 12/2013	2014	43	1.0	1.0	1.0	1.0	0.5	4.5	13	0
217	[42]	1997-2011	2015	86	1.0	1.0	0.5	0.0	1.0	3.5	2	0
219	[43]	to 2012	2015	99	1.0	0.5	1.0	1.0	1.0	4.5	2	0
222	[44]	1990-2012	2015	37	1.0	1.0	0.5	1.0	0.5	4.0	2	1
228	[45]	1997-1/2008	2015	29	1.0	0.5	0.0	1.0	0.0	2.5	0	1
236	[46]	2001-2013	2015	45	1.0	0.5	1.0	1.0	1.0	4.5	5	2
239	[47]	to 2011	2015	81	1.0	1.0	0.5	1.0	1.0	4.5	3	0
241	[48]	1980-2012	2015	87	1.0	1.0	1.0	1.0	0.5	4.5	6	0
244	[49]	1990-2012	2015	62	1.0	1.0	1.0	1.0	1.0	5.0	10	0
246	[50]	2003-4/2013	2015	20	1.0	1.0	1.0	0.0	0.5	3.5	2	1
249	[51]	2002-10/2012	2015	33	1.0	1.0	1.0	1.0	0.5	4.5	0	3
252	[52]	2002-2013	2015	30	1.0	0.0	0.5	1.0	1.0	3.5	6	0
259	[53]	1992-2/2014	2015	119	1.0	1.0	0.5	1.0	0.5	4.0	3	0
260	[54]	to 5/2015	2015	43(66)	1.0	0.5	0.5	0.0	0.5	2.5	10	0
268	[55]	1996-2/2008	2010	118	1.0	0.5	0.0	0.0	0.5	2.0	1	0
276	[56]	1996-10/2013	2015	31	1.0	1.0	0.5	0.0	0.5	3.0	2	0

ACCEPTED MANUSCRIPT

source journal, showing that all five journals did provide systematic reviews that were used in the final analysis. (The abbreviated journal titles from Table 2 have been used.)

Table 5: Number of systematic reviews at each stage

Journal	Found	Stage 2	Stage 3
EMSE	10	7	4
IST	97	42	21
JSS	31	14	7
SPE	5	3	1
TSE	13	8	4
Other	2	0	0
Totals	158	74	37

The additional data extraction performed specifically for STS1 was undertaken by two of us (DB and PB), on the basis that they were more familiar with the technical issues involved with synthesis. Again, this involved independent extraction, followed by the resolution of any differences.

5. Lessons about Reporting

The focus of this paper is upon the *reporting* of systematic reviews, drawing upon our experiences of performing part of the data extraction required for ETS2. There are two main reasons why we consider this to be an important issue for authors, journal referees, and readers.

1. Anyone planning to use the information from a review to guide practice or teaching needs to know about the *provenance* of any conclusions or recommendations in order to assess how appropriate it would be to adopt them in their own context.
2. A systematic review provides a 'snapshot' of empirical knowledge about a given topic at a particular point in time. Other researchers may seek to extend or augment such a review at a later date, and so will need the best possible information about how the review was performed.

Both have implications for the way that the review process is reported and how the outcomes are presented.

Table 6 shows the number of papers given each DARE rating. These profiles provide a useful indication of how thoroughly each part of the review process was performed, although they do need to be interpreted using the descriptions provided in Table 3. We should also point out that the DARE criteria address what should be reported rather than how it should be reported. And even then, they do not attempt to cover all issues, especially those that are not directly related to how the study is performed. For example, while we might expect any secondary study to include an assessment of threats to validity, this is not actually something identified as being a part of the DARE criteria.

Table 6: Profile of DARE score values

Score	incl./excl.	search	qual	desc.	syn
1.0	36	23	21	23	14
0.5	1	12	9	5	19
0.0	0	2	7	9	4

Within this section we therefore begin by examining some of the issues associated with each of the DARE criteria, and the associated lessons for reporting. We then examine the ways that outcomes are presented and review our experiences with seeking to identify *conclusions* and *recommendations*. Finally, we consider the limitations of this study.

5.1. DARE: inclusion-exclusion

As indicated in Table 6, this is the one DARE criterion where all studies scored more than zero, and indeed, most were considered to meet this fully.

Despite this, the inclusion/exclusion criteria themselves are not always clearly described (although many papers do identify them specifically). They are most easily recognised when listed as a table or in a bullet list. It is worth noting that all that is required to meet the DARE criterion is for the inclusion/exclusion criteria to be identified, and that it is not concerned with their clarity. We would also note that both inclusion and exclusion should be addressed.

Lesson 1: The inclusion/exclusion rules should be presented as a distinct element, such as a table, so that they can be readily recognised and cross-referenced.

As an example of this we can point to our own use of this form in Table 1.

The process involved in applying the inclusion/exclusion criteria is not always very clearly reported. In particular, it was not always clear how many people were involved in assessing each candidate paper, or even how this was organised. Reporting this is important, as from a quality perspective, the reader needs to know how reliable the assessments are likely to be.

Table 7 shows that many studies did use two reviewers, who then resolved any differences, but a substantial number still used a single reviewer with a checker. For many of these, the checker only checked a percentage of the selections (the lowest proportion of checks observed

ACCEPTED MANUSCRIPT

was 5% which is very weak). Five papers didn't report how this was done at all. One of the two papers described as 'other' had a quite complex multiple reviewer structure that was reported very clearly as a table [20], while in the other, the process of inclusion/exclusion appeared to be performed by two people working together, rather as in pair programming [41].

Table 7: Procedure used for inclusion/exclusion

Form used	No. studies
Two or more reviewers make independent decisions and resolve any differences	14
Two reviewers make independent decisions, and a third reviewer acts as an adjudicator for any differences	1
Two or more reviewers perform inclusion/exclusion, but not clear how this was organised	3
One reviewer makes decisions and another checks	11
Other	2
Not reported	6
Total	37

An assessment made by 'pooling' the independent results from two or three reviewers is likely to be more reliable than if only one reviewer has performed this task, with perhaps some checking by another author . As noted earlier, the use of multiple assessments is now recommended [10].

These descriptions were also scattered around different parts of the reports. Some were in the descriptions of 'planning', others in the descriptions of how a study was conducted, and a few could only be identified from the discussions of threats to validity or limitations.

Lesson 2: The number of reviewers performing each inclusion/exclusion decision should be reported as part of the description of how a study was conducted, and the mechanism for resolving differences arising among multiple reviewers should be described.

Our own processes were reported in Section 4.1.

We observed that the overall process was most usefully summarised as a diagram, giving the counts for the number of papers remaining at the different stages.

Lesson 3: The process of applying the inclusion/exclusion rules in order to produce the final body of evidence should be reported as a diagram, showing the different search elements, and the number of papers remaining at each stage.

For this study, this is provided by Figure 2.

5.2. *DARE: searching*

Table 6 shows that slightly fewer than two thirds of the studies were considered to have performed a search with good coverage as defined in Table 3. We formulated our interpretation of this criterion largely in terms of using electronic searching, which appears to be the normal approach adopted in software engineering. However, as with this study, there may be good arguments in favour of using a more focused and non-automated strategy and this does need to be kept in mind when interpreting the criterion.

The range of search engines used was quite wide, as indicated in Table 8, which shows the frequency with which each search engine was used (we have only included those used in five or more of the reviews).

Table 8: Search engines used

Search Engine	Number of uses
IEEExplore	33
ACM	30
ScienceDirect	24
Web of Science	19
Google Scholar	18
SpringerLink	17
Scopus	13
Compendex	11
CiteSeer	5

One observation is that *CiteSeer* appears to be losing popularity with less use of this in more recent studies. Another is the surprising number of studies that used *Google Scholar*, including one study that used it as the only search engine.

Table 9 shows the profile for the number of search engines used in the studies. The maximum number of search engines employed was 11, with the minimum being 0, the median value 5 and the mean value 5.4.

Many studies report the number of papers that were found per search engine, and a number also report the degree to which later searches found duplicates of those candidates already identified. In general, electronic searching was well reported. However, additional searching activities were less thoroughly reported, particularly where the use of manual searches of journals or conferences were concerned, and in particular, where any form of *snowballing* was employed. We noted one study where, because snowballing had not returned any additional papers, it was not included in the report on searching, and we were only able to identify that snowballing had been used because of a passing mention elsewhere in the report.

ACCEPTED MANUSCRIPT

Table 9: Number of search engines per study

No. search engines	No. studies
0	1
1	2
2	1
3	1
4	7
5	7
6	8
7	5
8	2
9	1
10	1
11	1

Lesson 4: All forms of searching together with the rationale for their use, and the numbers found for each form should be clearly reported, including nil returns. A recommended option is to include the numbers in the process diagram (see Lesson 3).

A noticeable feature in Table 4 is the wide variation in the way that the period covered by a systematic review was reported. In a few cases, we were completely unable to determine what this was, and for many others the information was incomplete, particularly regarding the start date. An open start date is of course acceptable, but this should be stated explicitly. Very few studies reported the date when searching took place, although knowing this can be important for anyone wanting to extend a review, or replicate it in some form.

Indeed, knowing both the complete value for the end date (e.g. 31st December 2015) and the date when the search was performed gives some idea about completeness. Digital library indexing is not always up to date (nor is the indexing of journals), so to be fairly sure of including all relevant studies published within a given period, it is prudent to conduct the search some time after the end date. Three months would seem to be a reasonable period to allow for this. This issue is relevant for manual searches as well of course.

Lesson 5: The exact period covered by the search and the date(s) on which electronic searches were conducted should be reported. (We suggest that this is again suitable for being presented as a small block or table.)

A related issue is the question of what constitutes a publication date for journal papers. Many journals now maintain an on-line list of papers 'in press' and report the date when a paper became available in this as part of the final publication. For this study, we regarded the publication date for a systematic review as being the date of the journal issue in which it was finally published. We suggest that systematic reviews make clear what their policy is with respect to 'in press' items, since these will often be found by electronic searches. In our protocol we did not explicitly specify that the publication date was to be treated as the date when a systematic review was assigned to a journal issue as one of our inclusion criteria, although in practice it formed one and is reported as such, since we did not include papers that were 'in press' during 2015.

Lesson 6: The inclusion criteria should make clear how the review will treat papers that are 'in press' at the time of an electronic search.

5.3. DARE: quality assessment

Table 6 shows that many of the systematic reviews did perform quality assessments for the primary studies, and most studies that did so, reported the questions they used and how they were derived. However, the DARE criterion is only concerned with whether quality scores were *derived* and not with whether they were *used*. Few of the reviews provided much detail about the quality scores for the primary studies, and there were relatively few examples of the quality score being used in any way during synthesis (or used at all).

This is illustrated in Table 10 where we examine the different ways in which the quality scores were used (if at all). In some cases very little detail about this was provided and there were sometimes statements about use that could not be substantiated from the available data.

Table 10: Use of quality scores in the selected studies

Form of use	Studies	Total
No quality scoring was performed	121, 123, 135, 154, 160, 193, 217, 246, 260, 276	10
Quality scores were derived, but no evidence for use	161, 167, 174, 175, 219, 228, 239, 249, 252	9
Quality scores were used for study selection	124, 146, 150, 155, 222, 236, 241, 244, 259, 268	10
Quality scores were used during analysis and synthesis	126, 130, 134, 138, 157, 197, 205, 215	8

Again, the associated processes were rarely described adequately, if at all. In this case there are two relevant processes that should be reported.

ACCEPTED MANUSCRIPT

- The process by which the quality questions were derived, such as whether or not they had been used in other studies, or were derived from other sets of questions. Quality questions do need to be relevant to the issues being addressed in the systematic review, and so need to be justified in some way.

- The process used to derive quality scores, including how many people performed each assessment and what mechanism was used to resolve any differences where there was more than one person performing the task.

Table 11 shows how the task of making a quality assessment was organised for the 27 papers that did perform a quality assessment. It is notable that this was less well reported than the procedures used to determine inclusion/exclusion.

Table 11: Procedure used for quality assessment

Form used	No. studies
Two or more reviewers make independent decisions and resolve any differences	8
Two reviewers make independent decisions, and a third reviewer acts as an adjudicator for any differences	3
Two or more reviewers make decisions, but it is not clear how this was organised	1
One reviewer makes decisions and another checks	6
Other	2
Not reported	7
Total	27

Lesson 7: Systematic reviewers should explain how quality questions were selected, and how the quality scores were derived, including the way that any difference in scores produced by using multiple reviewers were resolved.

For this study, the quality questions were provided by DARE and we have described the process of performing quality assessment in Section 3.3.

As noted in [10] the purpose of using quality scores is to enhance a systematic review, for example by weighting the importance of individual primary studies when determining study outcomes, or by guiding the way that the outcomes are interpreted. Table 10 shows little to indicate that this is common practice in software engineering, with only eight studies from 37 using the quality analysis in this way, and half of the studies (19) either performing no scoring of quality or performing one and not using it. This leaves the question open as to why researchers performed a quality analysis and then made little use of it, other than because it was recommended in the guidelines.

A number of studies (10) used the quality scores as part of the selection process, usually by omitting those primary studies that had scores below some (arbitrary) threshold. Not all of them reported much about the studies that were discarded or about the reasons for choosing a particular threshold value.

One concern about this practice is that the choice of a threshold value introduces an non-systematic element into the selection process. It also muddles the issue of quality assessment with inclusion/exclusion rules. Overall, it seems undesirable to conflate quality assessment with selection in this way.

One reason for the profile shown in Table 10 may be that it is linked to a lack of confidence about the process of synthesis, a point that we will return to later.

Lesson 8: Systematic reviewers should explain why they are performing a quality assessment of the primary studies (or otherwise) and the role of quality assessment should be kept clearly distinct from the process of study selection.

5.4. DARE: study descriptions

There are examples of this that range from providing hardly any information about the primary studies up to the provision of quite detailed information using tables. Some use one table effectively, others use multiple tables effectively and there is probably no one clear lesson here. Most give bibliographic information. However, many studies did score zero for this one. This was largely because systematic reviews often provided little information about the primary studies and their characteristics (providing only bibliographic information would lead to a score of 0).

Useful information about the primary studies can be considered as anything that is related to the issue of *provenance*. While meta-data such as date of publication, location of authors etc. does not help with this, details of the study itself can provide an understanding of the process of synthesis (addressed in the next subsection).

What comprises useful information will depend to some extent upon the topic of the systematic review and its research questions, but is likely to include some or all of the following.

- The *form* of the primary study, such as an experiment, case study etc.

ACCEPTED MANUSCRIPT

- The *context* for the study (where conducted and by whom, whether a replication etc.).
- The *size* of the context, such as the number of participants or the size of a system.
- The *type* of the participants, for example, undergraduate students, practitioners with more than 5 years experience.
- The *source* of any material used in the study, such as student projects, industrial ones, open source etc.

Two good examples of very different studies that provide the sort of detail indicated above are [32] (Tables 2 and 3), and [47] (in Appendix B.)

As Table 4 shows, providing such a profile for the studies can make it possible to look at the data-set as a whole, and to spot factors that might be of interest.

A reporting issue that was encountered with a number of the systematic reviews was that the authors failed to make clear when they were counting *papers* and when they were counting *studies*. Empirical papers quite commonly report the results of more than one study, and for a secondary study it is usually appropriate to treat these as individual inputs. This complicates the reporting of counts, since for searching and inclusion/exclusion the relevant operational unit is the paper, while for analysis it is the study. (And of course, there is the added complication that conference papers may be extended for journal publication, making the relationship between papers and studies to be many-to-many.) And as a further complication, a review might include one study from a paper, while excluding another.

Since studies are the 'atomic' unit of measurement when analysing and interpreting the outcomes of the primary studies, we would advocate that all counts related to analysis should make reference to studies. Where necessary, the reporting of the included studies should make clear the relationship between papers and studies.

Lesson 9: All counts reported in the analysis and interpretation of a systematic review should relate to individual studies, not papers.

5.5. DARE: synthesis

As we were seeking systematic reviews rather than mapping studies, it is perhaps not surprising that few reviews scored zero under this heading, although there were a lot of half-scores. Many papers don't make their synthesis method clear (including misleading titles about mapping studies) but there are good examples of ones that do give counts and details of any papers that support their observations.

For the purpose of this paper we used the set of definitions for forms of synthesis provided in Table 2 of the tertiary study performed by Cruzes & Dybå [15], cross-checking our assessments against theirs where possible. As noted earlier, the task of classification was additional to the data extraction performed for ETS2, and was undertaken by only two of us (DB and PB), as it was felt that this would provide greater consistency of interpretation.

Our analysis of synthesis was complicated by a number of issues. One is that many authors do not describe the form of synthesis employed, or if they do, they may have used terms taken from other sources (and the descriptions are not always even consistent across different sections). A second is that more than one form of synthesis may well be used within a systematic review to answer the different research questions. A third complication is that sometimes, while there has been an element of synthesis, this may be related to how the primary studies performed their research rather than the outcomes. We have summarised our findings in Table 12, using the following coding conventions.

- Where the study reference number is in italics this indicates that this was the form of synthesis identified by both the authors of the review and also by ourselves. Otherwise, the classification is one that we have assigned for the study.
- Where the study reference is in parentheses, this indicates that more than one form of synthesis was employed in the study.
- Where the study reference is in square brackets, this indicates that any synthesis performed related to primary study forms rather than outcomes.

Most systematic reviews did perform some form of synthesis, as might be expected, given the criteria used to select them, although in a few cases we did consider that the outcomes were useful for other reasons.

As a consistency check on our coding of synthesis forms, we 'blinded' ourselves to the outcomes of the study in [15] until we had completed coding. We then looked to see how many systematic reviews were included in both that tertiary study and this one. Table 13 summarises how our codings compared with those of Cruzes & Dybå (labelled as 'C&D') for the six studies that were common to both. We provide the study index values that were used in both tertiary studies.

Closer examination of the one paper that showed a significant difference (124) revealed that it had a rather

ACCEPTED MANUSCRIPT

Table 12: Forms of synthesis used

Form of synthesis	Studies	Total
None identifiable	123, [167], [174] [259]	4
Meta-Analysis	*157*, *217*	2
Narrative	*121*, (*130*), *138*, 146, 150, 154, 193, *244*, 268	9
Grounded Theory	*246*	1
Thematic Analysis	124, 135, 155, 160, 161, 175, 215, 228, (*236*), 239, (*241*), 252, 260, (276)	14
Vote Counting	*126*, (*130*), *134*, 197, *205*, *219*, (*236*), (241), (276)	9
Case Survey	*222*	1
Content Analysis	249	1

Table 13: Synthesis coding for common studies

Our index	Our Coding	C&D index	C&D Coding
121	Narrative	S31	Narrative
123	None	S22	Not explicit/ scoping
124	Narrative	S30	Thematic
126	Vote Counting	S46	Comparative Analysis using vote counting
130	Narrative + Vote Counting	S45	Narrative
268	Narrative/ Scoping	S33	Scoping

unusual structure, with the outcomes of synthesis being used to build a model. This model (for a change characterisation scheme) could therefore be interpreted as a thematic analysis, and we decided to alter our interpretation to be consistent with that of Cruzes & Dybå, leading to the value presented in Table 12.

It is significant that only 13 of the 37 papers described the form(s) of synthesis employed. Inevitably perhaps, there is an overlap between these and the set of studies that used the quality scores during synthesis (130, 134, 138, 157, 205).

Lesson 10: The form(s) of synthesis used for the research question(s) should be reported. Where possible, the quality scores should be used as part of synthesis.

5.6. Identifying the outcomes

One of the differences between ETS1 and ETS2 is that the former study did not attempt to make detailed assessments of the nature and quality of knowledge provided in a systematic review, nor of its *provenance*. For ETS2, we distinguished between conclusions and recommendations, and also sought to extract data related to the confidence that could be placed upon these.

Even having made this distinction, this was one element of data extraction where we often struggled to identify the relevant information, and where disagreement between team members did occur quite frequently. There are two clear reasons why this is so:

- The relevant information is apt to be spread among the later sections of a paper, necessitating thorough scrutiny of 'discussion' sections as well as 'conclusions', and sometimes the 'results' sections too. Only a few papers provided tabulated presentations of results that showed which primary studies supported or refuted a particular conclusion.
- Provenance in the form of a link between a conclusion (or recommendation) and the supporting data is often unclear. Since we were only prepared to include those conclusions or recommendations that were explicitly supported by the primary studies, this lack of clear links often made it quite difficult to identify where such support existed.

Both of these factors could partly be remedied by better reporting, although arguably, many systematic reviews also need better analysis and synthesis. Even where papers did report which studies *supported* a conclusion or recommendation, they sometimes failed to identify any studies that did *not* support it, although they may well have existed. Tables 5-8 in [40] provide a good example of how such information can be presented.

We noted a relative lack of *recommendations*. Since identifying these is essentially the task of *knowledge translation*, and this process is still ill-defined even for those disciplines with a longer tradition of using systematic reviews, this should not be that surprising. Identifying recommendations does also require domain expertise, and it may well be that many of the systematic review teams did not feel confident to do so. (It might be argued that it is better not to do so than to do it badly!)

Lesson 11: The key findings of a study should be clearly reported. These should be summarised in a block (or table) so that they can be easily identified by users, together with information about their provenance.

Lesson 12: Findings should be reported as 'conclusions' unless there is strong evidence, combined with

ACCEPTED MANUSCRIPT

domain expertise, that can be used to formulate 'recommendations', which should include an indication of their strength.

5.7. Limitations

There are some limitations that are implicit in the way that we conducted our analysis for STS1.

- Our study selection process. We did not attempt to find all of the systematic reviews published in this period, limiting our selection to five major software engineering journals. Also, since our quality assessment is based upon systematic reviews that contained material relevant to education and practice, it may not reflect the way that more research-oriented studies were reported. However, in performing the tasks of inclusion/exclusion we did examine many other studies that were later excluded, and did not observe any significant differences in the way that these were reported.
- Data extraction. Most of the information extracted was objective, such as details about the body of knowledge used in a systematic review, the process followed, and the use of quality assessment. For this, the main risk was that of missing something that was presented in a non-standard manner, and we would consider this to be relatively low. However were two element of data extraction that involved some element of interpretation, and for which we tried to maintain a high degree of rigour.
  - The form(s) of *synthesis* employed in a paper. We tried to ensure consistency of interpretation by using the same two analysts for this element, and sought to minimise bias by checking against the interpretations provided by Cruzes & Dybå [15].
  - Identification of *conclusions* and *recommendations*. Few papers identified these clearly or explicitly, and so this did require that we examined all of the later sections of each paper using two analysts and discussing the results in detail. As a further check, we consulted the original authors wherever possible, and almost all responses concurred with our extracted outcomes.
- Derivation of the *Lessons*. These are largely identified in relation to the main elements of our analysis. They seek to capture our collective observations about the systematic reviews and hence have been discussed by the team. However, they were not derived through the use of any form of systematic process.

6. Conclusions

Our assessment of the ways that the 37 systematic reviews were reported identifies both good and undesirable aspects of both the reporting process, and also by implication, of the manner in which systematic reviews are currently being employed in software engineering. There is good evidence that many researchers are performing thorough searches through the literature, and that they are using rigorous inclusion/exclusion procedures to select the relevant primary studies.

However, we observed that not only do few systematic reviews in software engineering provide material that is likely to be useful for teaching or for practitioners, but even when such material is available, they do not report it in a clear and effective manner. There is also clear evidence that quality assessment of the primary studies is not used consistently, or sometimes not used at all. In part this may arise because the forms of synthesis used are often unsuited to the use of quality weightings (narrative synthesis in particular). Where we did observe the use of synthesis in these studies there was little use of more quantitative approaches such as vote counting.

In this paper we have concentrated on reporting what was found, and how it might be improved (the role of the *Lessons*), and have not attempted to identify the causes for what we have observed. The use of systematic reviews in software engineering is relatively new, although this has clearly been quite widely adopted. As such it is therefore a good time to look at how this use is being adapted to the needs of software engineering and identify ways to improve this, as we have sought to do, with the *Lessons* encapsulating our findings.

It would appear that even when we use sound procedures, poor reporting may mean that the discipline of software engineering is not obtaining the best value from the use of systematic reviews. Both synthesis and reporting could be improved and the outcomes made more useful to practitioners and teachers. Based upon our *Lessons* we have identified a checklist for journal and conference referees (and authors) in Appendix A and suggest that adopting (and refining) this can provide a practical step towards encouraging better practice when reporting systematic reviews.

ACCEPTED MANUSCRIPT

Appendix A. Checklist for Authors and Referees

The refereeing process employed by journals and conferences provides an important element of 'quality control' for any discipline. Based upon the issues related to reporting of systematic reviews identified in this paper, we suggest that referees be encouraged to ensure that accepted papers provide at least the information summarised in Table A.14. And of course, if the check-list is relevant for referees then it should also be useful for authors.

Acknowledgements

Our thanks to Professor Barbara Kitchenham for advice and observations, as well as for conducting independent electronic searching for publications that we might have missed. Our thanks also to the authors of the systematic reviews we studied, many of whom were good enough to check the accuracy of our extracted conclusions and to pass comment on these. And finally, our thanks to the anonymous reviewers for a number of helpful suggestions about both content and presentation of this paper.

References

[1] B. Kitchenham, T. Dybå, M. Jørgensen, Evidence-based software engineering, in: Proceedings of ICSE 2004, IEEE Computer Society Press, 2004, pp. 273–281.

[2] S. Oliver, K. Dickson, Policy-relevant systematic reviews to strengthen health systems: models and mechanisms to support their production, Evidence & Policy 12 (2016) 235–259.

[3] K. Petersen, S. Vakkalanka, L. Kuzniarz, Guidelines for conducting systematic mapping studies in software engineering: An update, Information & Software Technology 64 (2015) 1–18.

[4] B. Kitchenham, P. Brereton, D. Budgen, Mapping study completeness and reliability—a case study, in: Proceedings of 16th EASE Conference, IET Press, 2012, pp. 1–10.

[5] B. Kitchenham, P. Brereton, D. Budgen, The educational value of mapping studies of software engineering literature, in: Proceedings ICSE'10, ACM Press, 2010.

[6] D. Budgen, S. Drummond, P. Brereton, N. Holland, What scope is there for adopting evidence-informed teaching in software engineering?, in: Proceedings of 34th International Conference on Software Engineering (ICSE 2012), IEEE Computer Society Press, 2012, pp. 1205–1214.

[7] D. Budgen, P. Brereton, N. Williams, S. Drummond, A tertiary review of evidence about software engineering practice, 2017. Paper in preparation.

[8] P. Brereton, B. A. Kitchenham, D. Budgen, M. Turner, M. Khalil, Lessons from applying the systematic literature review process within the software engineering domain, Journal of Systems and Software 80 (2007) 571 – 583.

[9] B. Kitchenham, S. Charters, Guidelines for performing systematic literature reviews in software engineering, Technical Report, Keele University and Durham University Joint Report, 2007.

[10] B. A. Kitchenham, D. Budgen, P. Brereton, Evidence-Based Software Engineering and Systematic Reviews, Innovations in Software Engineering and Software Development, CRC Press, 2015.

[11] B. A. Kitchenham, D. Budgen, O. P. Brereton, Using mapping studies as the basis for further research—a participant-observer case study, Information & Software Technology 53 (2011) 638–651. Special section from EASE 2010.

[12] B. Kitchenham, P. Brereton, D. Budgen, M. Turner, J. Bailey, S. Linkman, Systematic literature reviews in software engineering — a systematic literature review, Information & Software Technology 51 (2009) 7–15.

[13] B. Kitchenham, R. Pretorius, D. Budgen, P. Brereton, M. Turner, M. Niazi, S. Linkman, Systematic literature reviews in software engineering — a tertiary study, Information & Software Technology 52 (2010) 792–805.

[14] F. Q. da Silva, A. L. Santos, S. Soares, A. C. C. França, C. V. Monteiro, F. F. Maciel, Six years of systematic literature reviews in software engineering: An updated tertiary study, Information and Software Technology 53 (2011) 899–913.

[15] D. S. Cruzes, T. Dybå, Research synthesis in software engineering: A tertiary study, Information and Software Technology 53 (2011) 440 – 455.

[16] D. Budgen, B. Kitchenham, P. Brereton, The Case for Knowledge Translation, in: Proceedings of 2013 International Symposium on Empirical Software Engineering & Measurement, IEEE Computer Society Press, 2013, pp. 263–266.

[17] K. S. Khan, R. Kunz, J. Kleijnen, G. Antes, Systematic Reviews to Support Evidence-Based Medicine, Hodder Arnold, 2nd edition, 2011.

[18] M. Ardis, D. Budgen, G. W. Hislop, J. Offutt, M. Sebern, W. Visser, SE2014: Curriculum Guidelines for undergraduate degree programs in software engineering, IEEE Computer (2015) 106–109.

[19] B. Kitchenham, P. Brereton, A systematic review of systematic review process research in software engineering, Information and Software Technology 55 (2013) 2049 – 2075.

[20] D. Smite, C. Wohlin, T. Gorschek, R. Feldt, Empirical evidence in global software engineering: a systematic review, Empirical Software Engineering 15 (2010) 91–118.

[21] L. B. Lisboa, V. C. Garcia, D. Lucrédio, E. S. de Almeida, S. R. de Lemos Meira, R. P. de Mattos Fortes, A systematic review of domain analysis tools, Information and Software Technology 52 (2010) 1 – 13.

[22] B. J. Williams, J. C. Carver, Characterizing software architecture changes: A systematic review, Information & Software Technology 52 (2010) 31–51.

[23] M. Turner, B. Kitchenham, P. Brereton, S. Charters, D. Budgen, Does the technology acceptance model predict actual use? A systematic literature review, Information and Software Technology 52 (2010) 463 – 479.

[24] M. S. Ali, M. A. Babar, L. Chen, K.-J. Stol, A systematic review of comparative evidence of aspect-oriented programming, Information and Software Technology 52 (2010) 871 – 887.

[25] O. Dieste, N. Juristo, Systematic review and aggregation of empirical studies on elicitation techniques, IEEE Transactions on Software Engineering 37 (2011) 283–304.

[26] A. H. Ghapanchi, A. Aurum, Antecedents to IT personnel's intentions to leave: A systematic literature review, Journal of Systems & Software 84 (2011) 238–249.

[27] K. Peterson, Measuring and predicting software productivity: A systematic map and review, Information & Software Technology 53 (2011) 317–343.

[28] T. B. C. Arias, P. van der Spek, P. Avgeriou, A practice-driven systematic review of dependency analysis solutions, Empirical

ACCEPTED MANUSCRIPT

Software Engineering 16 (2011) 544–586.

[29] J. Díaz, J. Pérez, P. P. Alarcón, J. Garbajosa, Agile product line engineering–a systematic literature review, Software — Practice and Experience 41 (2011) 921–941.

[30] C. Zhang, D. Budgen, What do we know about the effectiveness of software design patterns?, IEEE Transactions on Software Engineering 38 (2012) 1213–1231.

[31] T. Hall, S. Beecham, D. Bowes, D. Gray, S. Counsell, A systematic literature review on fault prediction performance in software engineering, IEEE Transactions on Software Engineering 38 (2012) 1276–1304.

[32] Y. Rafique, V. Misic, The effects of test-driven development on external quality and productivity: A meta-analysis, IEEE Transactions on Software Engineering 39 (2013).

[33] A. M. Magdaleno, C. M. L. Werner, R. M. de Araujo, Reconciling software development models: a quasi-systematic review, Journal of Systems & Software 85 (2012) 351–369.

[34] C. Pacheco, I. Garcia, A systematic literature review of stakeholder identification methods in requirements elicitation, Journal of Systems & Software 85 (2012) 2171–2181.

[35] Z. Li, H. Zhang, L. O'Brien, R. Cai, S. Flint, On evaluating commercial cloud services: A systematic review, Journal of Systems & Software 86 (2013) 2371–2393.

[36] N. B. Ali, K. Peterson, C. Wohlin, A systematic literature review on the industrial use of software process simulation, Journal of Systems & Software 97 (2014) 65–85.

[37] S. U. Khan, M. Niazi, R. Ahmad, Barriers in the selection of offshore software development oursourcing vendors: An exploratory study using a systematic literature review, Information & Software Technology 53 (2011) 693–706.

[38] R. Giuffrida, Y. Dittrich, Empirical studies on the use of social software in global software development–A systematic mapping study, Information & Software Technology 55 (2013) 1143–1164.

[39] D. Radjenović, M. Heričko, R. Torkar, A. Živkovič, Software fault prediction metrics: A systematic literature review, Information & Software Technology 55 (2013) 1397–1418.

[40] H. Munir, M. Moayyed, K. Peterson, Considering rigor and relevance when evaluating test driven development: A systematic review, Information & Software Technology 56 (2014) 375–394.

[41] N. Paternoster, C. Giardino, M. Unterkalmsteiner, T. Gorschek, Software development in startup companies: A systematic mapping study, Information & Software Technology 56 (2014) 1200–1218.

[42] U. Abelein, B. Paech, Understanding the influence of user participation and involvement on system success — a systematic mapping study, Empirical Software Engineering 20 (2015) 28–81.

[43] R. Jabangwe, J. Borstler, D. Smite, C. Wohlin, Empirical evidence on the link between object-oriented measures and external quality attributes: a systematic literature review, Empirical Software Engineering 20 (2015) 640–693.

[44] O. Al-Baik, J. Miller, The Kanban approach between agility and leanness: a systematic review, Empirical Software Engineering 20 (2015) 1861–1897.

[45] M. Zarour, A. Abran, J.-M. Desharnais, A. Alarifi, An investigation into the best practices for the successful design and implementation of lightweight software process assessment methods: A systematic literature review, Journal of Systems & Software 101 (2015) 180–192.

[46] A. Nguyen-Duc, D. S. Cruzes, R. Conradi, The impact of global dispersion on coordination, team performance and software quality – a systematic literature review, Information & Software Technology 57 (2015) 277–294.

[47] F. S. Silva, F. S. F. Soares, A. L. Peres, I. M. de Azevedo, A. P. L. F. Vasconcelos, F. K. Kamei, S. R. de Lemos Meira, Using CMMI together with agile software development: A systematic review, Information & Software Technology 58 (2015).

[48] M. Bano, D. Zowghi, A systematic review on the relationship between user involvement and system success, Information & Software Technology 58 (2015).

[49] A. Idri, F. A. Amazal, A. Abran, Analogy-based software development effort estimation: A systematic mapping and review, Information & Software Technology 58 (2015) 206–230.

[50] I. Steinmacher, M. A. G. Silva, M. A. Gerosa, D. F. Redmiles, A systematic literature review on the barriers faced by newcomers to open source software projects, Information & Software Technology 59 (2015).

[51] M. Brhel, H. Meth, A. Maedche, K. Werder, Exploring principles of user-centered agile software development: A literature review, Information & Software Technology 61 (2015) 163–181.

[52] E. Kupiainen, M. V. Mäntylä, J. Itkonen, Using metrics in agile and lean software development – a systematic literature review of industrial studies, Information & Software Technology 62 (2015) 143–163.

[53] S. Tiwari, A. Gupta, A systematic literature review of use case specifications research, Information & Software Technology 67 (2015) 128–158.

[54] D. Heaton, J. C. Carver, Claims about the use of software engineering practices in science: A systematic literature review, Information & Software Technology 67 (2015) 207–219.

[55] R. Rabiser, P. Grunbacher, D. Dhungana, Requirements for product derivation support: Results from a systematic literature review and an expert survey, Information & Software Technology 52 (2010) 324–346.

[56] E. Tüzün, B. Tekinerdogan, M. E. Kalender, S. Bilgen, Empirical evaluation of a decision support model for adopting software product line engineering, Information & Software Technology 60 (2015) 77–101.

ACCEPTED MANUSCRIPT

Table A.14: Essential Information that should be reported about a Systematic Review: Referee (and Author) Checklist

Review aspect	Information Required	Rationale
Inclusion/Exclusion	The rules for both inclusion and exclusion should be clearly stated.	This information is important for understanding the scope of a systematic review.
Inclusion/Exclusion	How the rules were applied and any differences between reviewers were resolved should be described.	This is a quality issue that should provide confidence in the procedures used to perform the review.
Inclusion/Exclusion	The number of papers remaining at each stage of selection should be reported.	This forms a part of the provenance for the study itself.
Searching	All of the search mechanisms used should be clearly reported.	This again relates to provenance, and the reasons for choosing a particular search strategy should be made clear.
Searching	The period covered by the search should be explicitly stated, and the dates when any searches were performed should be reported.	This will aid any future systematic reviews that seek to extend the results.
Quality Assessment	When performed, the intended use as well as the checklist items should be reported.	Quality assessment is normally used to assist synthesis, and if used as part of selection this needs to be explained and justified.
Quality Assessment	How quality assessment was undertaken, and how any differences between reviewers were resolved need to be explained.	This will help provide confidence in the way that the review was performed.
Synthesis	Where performed, the form of synthesis adopted for specific research questions should be described, and the reason for its use should be given.	This is part of the information needed to demonstrate the provenance of any findings from the review.
Outcomes	Key findings should be clearly reported, together with any information related to the 'strength of evidence' that applies to them.	This is part of the information needed to establish the provenance of the findings and what confidence can be given to them.

ACCEPTED MANUSCRIPT

https://dspace.org/

Reporting systematic reviews: Some lessons from a tertiary study

Budgen, David; Brereton, Pearl; Drummond, Sarah; Williams, Nikki

2017-10-27
Elsevier
Budgen D, Brereton P, Drummond S, Williams N. (2017) Reporting systematic reviews: Some lessons from a tertiary study. Information and Software Technology, Volume 95, March 2018, pp. 62-74

Downloaded from a DSpace repository. https://dspace.org/