Reviewing the Systematic Review

William Wallis — 2025138W

December 20, 2016

Abstract

Systematic reviewing is a technique for bringing scientific rigour to a computer science literature review, pioneered by Barbara Kitchenham (Kitchenham, 2004). 12 years after Kitchenham's original guidelines were set for structuring a systematic literature review, the technique has seen widespread adoption — but the original guidelines raise questions and note possible issues with the method. A review of these systematic reviews may highlight whether these concerns are worth revisiting, before Kitchenham's guidelines and those like them become standard practice for the software engineering research community.

1 Introduction

A Systematic Review is a literature review which collates the results of many papers, using statistical analysis to draw empirical results about the state of a research field and to answer research questions posited as the motivation for the review. Systematic reviews are born from the philosophy that a literature review should have scientific merit, and produce reproducible results. This technique somewhat opposes standard techniques for literature reviews, which leave more room for subjective insight. The scientific nature of a systematic literature review, in theory, removes ambiguity and bias from literature review practice and lends the review the same validity and credence as a research study, though a statistical analysis of collated results.

Kitchenham's systematic literature review technique[1] has begun to dominate as a literature review technique for software engineering. Kitchenham's review procedure stems largely from literature review techniques in medical research (Kitchenham, 2004; Khan et al., 2001), where empirical studies which verify the validity of literature already published in peer-reviewed journals is paramount to civilian safety. Conventional computing

[1]Kitchenham's guidelines have undergone some revisions over time — the two most cited versions in the papers reviewed being Kitchenham (2004) and Kitchenham and Charters (2007). As the latter is an incremental improvement over the former, all of Kitchenham's guideline versions will be referred to through this document simply as Kitchenham's guidelines.

science literature reviews might closer resemble Webster's guidelines (Webster and Watson, 2002), which are less rigorous, and less focused on empiricism and repeatability, yet offer structure to the review. However, doubts are sometimes raised. For example, in Kitchenham's own guidelines:

> In particular, software engineering research has relatively little empirical research compared with the large quantities of research available on medical issues, and research methods used by software engineers are not as rigorous as those used by medical researchers.

Kitchenham (2004)

Kitchenham does provide types of empirical data which software engineering research produces, which can be appropriate for statistical analysis in a literature review. However, whether research rigour and types of data collected are make appropriate note of by the research community is clear only now that a wealth of systematic literature reviews have been produced.

In this review, a series of systematic literature reviews will be analysed and searched for their scrutiny of research rigour and format of empirical data. In this way, the importance of this doubt regarding the suitability of systematic reviews for software engineering research will be assessed. We will see that there is some reason for this doubt, and offer solutions to the problem as it manifests.

The reviews chosen were picked as a result of their popularity on the "*Google Scholar*" academic search engine, found by a search for "software engineering "systematic" literature review", and similar searches. This was to find papers which were well-cited and high-impact, because as the question to be answered would impact the culture around systematic reviews, these papers are important, as they are most likely to influence future systematic reviews.

2 Papers reviewed

2.1 Effect Size in Software Engineering

First reviewed is a study into effect size in software engineering experiments: Kampenes, Dybå, Hannay, and Sjøberg (2007). The review was conducted according to systematic review guidelines — the guidelines were not specified, however the procedure roughly aligned with that of Kitchenham, and Kitchenham's guidelines were cited as reasoning when defining a search criteria.

As required when developing a systematic review, Kampenes et al. (2007) used search criteria to determine the literature which was to be reviewed. This search criteria was cited from another paper (Sjøberg et al. (2005), which shares some authors), rather than

stated directly; the collated work was a set of software engineering papers exhibiting controlled experiments published over the span of 10 years.

The method for data extraction was also well reported, as a systematic review should entail by Kitchenham's guidelines. The paper goes on to perform a deep and thorough statistical analysis, and concludes with a review of the results of these analyses with a comparison to the results of similar papers in Psychology and Behavioural Science. The paper succeeded in selecting several papers with empirical data so as to perform the statistical analysis with a large sample. The authors found 78 usable studies for their review, from the 5453 studies assessed. Happily, the researchers found quantitative data on which to perform some statistical analysis.

2.2 Global Software Engineering

Šmite, Wohlin, Gorschek, and Feldt (2010) sets about the task of reviewing literature on Global Software Engineering (GSE). Particularly, it attempts to collate and assess the results of literature which produce empirical data. The authors identify that there exists scarce literature on the topic, and so to collate the findings and categorise the growing yet important field, they employ a systematic review as a technique for categorising literature based on emerging trends.

The review guidelines used were from a Kitchenham-like standard (Kitchenham and Charters, 2007). The authors do not give a justification for the use of a systematic technique as opposed to a regular review. However, they do note that no systematic review yet existed — so one is inclined to suppose that the authors sought to fill the niche they had identified. Šmite et al. do present a useful section explaining their search method, useful for repetition of a systematic review. The authors found 387 papers which were possibly relevant, of which 59 were considered for their literature review.

As the field is young, this literature review serves to add to the literature present and to summarise the current state of the literature. It also contributes useful categories by which future research in the field might be defined — as a literature review surveys much of the original research, this serves as a significant contribution to the field, directing future work. Overall, the work serves to help in guiding GSE research by collating existing research into the categories defined.

2.3 Automated Analyses of Feature Models

Benavides, Segura, and Ruiz-Cortés (2010) performs a review of techniques used for automating the analysis of a feature model. The review is an interesting blend of Kitchenham's systematic approach and the more semantic (yet structured) approach of Webster's guidelines. In doing this, Benavides et al. produce a rigorous review which also persistently focuses on making its findings accessible to the reader, though data visualisation and detailing the non-quantitative findings of the review.

Benavides et al. detail their effective research method as well as three research questions, which are closely examined. They also detail their inclusion and exclusion criteria for their search, enabling the repeatability of the systematic review.

The authors conclude their research, having assessed the state-of-the-art in the field and observing the current research trends, with an exposition as to future challenges the field may face. Unfortunately, the majority of their results are visualisation and discussion as opposed to statistical analysis, as per Kitchenham's method. These visualisations are however created in such a way that their review may be repeated and similarities observed, meaning that some degree of repeatability is still present.

In this work, 53 primary studies were reviewed from a total of 72 candidates, an exceptionally high proportion of the original set compared to other papers reviewed here.

2.4 Motivation of Software Engineering

Beecham, Baddoo, Hall, Robinson, and Sharp (2007) reviews how motivation affects the process of software engineering. Specifically, the study assesses how software engineers might me mtivated or unmotivated, and what factors cause this to happen on a personal level. It also assesses what properties software engineers tend to have, and what models of motivation exist in software engineering.

This study followed the Kitchenham guidelines for conducting their review. With five research questions answered over the course of the paper, the research is involved and lengthy — however, no statistical analysis is produced. Instead, it uses representations of the collated research, alongside counts of traits of software engineers in the various studies. Kitchenham's guidelines are lauded in part for their rigour and repeatability. While this research contained both, the lack of data to analyse and semantic nature of the findings means that a repeatability study may not hold much value. Together with the many visual aids presented, this review may have benefited from techniques in Webster's method, which was was similar to in nature (yet systematic).

2.5 Variability Management in Software Product Lines

Chen, Babar, and Ali (2007) review the field of Variability Management according to Kitchenham's guidelines. While there have been other literature reviews in the field, this systematic review is the first to rigorously assess it.

Chen et al. select 34 papers from an original sample of 628. While the procedure to collect the original 628 is unclear, their inclusion and exclusion criteria from that point forward are well documented. In terms of adherence to systematic review guidelines, Chen et al. do not succeed in producing a model or developing statistical inferences from the data produced. However, they do perform a textual analysis to create a repeatable assessment of the current state-of-the-art in Variability Management.

The paper also summarises problems with current literature: only a handful of publications studied tackled important issues such as scalability and testing. No mention of inspection and review as quality assurance techniques were observed by the authors. In this way, while no statistical analysis was completed, the authors successfully assess the current state of the literature and provide many points of potential improvement for researchers in the area.

2.6 A Systematic Review of Systematic Reviews

A final paper worth noting for the purposes of this review is Kitchenham's own systematic review of published systematic reviews (Kitchenham and Brereton, 2013). The review by Kitchenham and Brereton is very thorough, and includes research questions, search criteria, and all other characteristic traits of a systematic review.

This review excluded papers to a pool of 45 from an original sample of 410. Like other reviewed papers, this review creates taxonomies and identifies key traits of a review through textual analysis rather than statistical analysis. However, the review also includes statistical analyses of these taxonomies, provided statistics such as Kappa calculations. It is the only paper reviewed to do so.

This paper selects papers with appropriate impacts and represents the state of the art in systematic reviews very effectively. In concluding the review, the authors highlight some concerns with current review procedure, as well a concerns with current guidelines and best practices, offering constructive criticism of changes to their own work and others in light of their findings. It also calls for further research into systematic reviews, and for better tooling to improve the issues they identify, which include ceasing to mention data extractors and checkers and to mention citation-based search strategies.

3 Discussion

The systematic reviews selected have all been reviews of literature pertaining to a niche in software engineering. Interestingly, a common theme emerged, which was that there was frequently scarce quantitative data for the reviewers to analyse statistically.

This seemed to occur for two reasons:

1. Some subject areas did not easily produce empirical quantitative data. An example of this would be the Benavides et al. (2010) article on Motivation; achieving consistently reliable quantitative data in psychological or ethnographic problem domains is difficult to do.

2. The research culture of some subject areas did not seem inclined to promote the production and publication of quantitative data as a part of their work. Examples of this would be Chen et al. (2007), or Šmite et al. (2010).

Unsurprisingly, Kitchenham and Brereton (2013) succeeds in creating statistical analyses from the taxonomic data the authors collate over the course of the review. Also worth noting is that another paper, **?**, *did* succeed in creating statistical measurements from the taxonomical data produced. One of the authors, however — Dyba — has their own systematic review procedure published (Dyba et al., 2005), so this comes at little surprise.

For systematic reviews to have a more scientific advantage over more orthodox review techniques, quantitative data evaluation is paramount. It remains the most reliable and clearest metric of a piece of scientific research's reproducibility.

For review culture where textual analysis is intended to go hand-in-hand with repeatability, a review method such as Webster and Watson (2002) may be more appropriate. Webster's method offers a helpful guide for getting away from the "phonebook" issues some literature reviews encounter. In addition, the repeatable nature of the taxonomical data they produce fits nicely in line with Webster's method.

An alternative method for solving these issues would be to adopt the statistical analysis methods from Kitchenham and Brereton (2013) or Chen et al. (2007). This would allow at least modest statistical analysis of taxonomical data, should researchers be particularly inclined toward the systematic philosophy of literature reviewing.

4 Conclusion

The literature reviewed here, while of quality and all of impact, did not consistently fulfil the criteria which provides value to a systematic review. Indeed, Kitchenham's early doubt — that software engineering might not produce enough quantitative data to bring scientific rigour to the review in this case — seems well-placed in this instance.

Many solutions are available to remedy this, however. Guidelines more in line with statistical analysis of taxonomical data is one option — of which Kitchenham and Brereton (2013) and **?** are good examples. Other options would be the adoption of structured review methods where systematic reviews may be less appropriate, for which Webster and Watson (2002) would be a good choice. A final option would be, as a culture, to shift more toward quantitative data production and analysis... though this seems impractical as a solution.

References

Sarah Beecham, Nathan Baddoo, Tracy Hall, Hugh Robinson, and Helen Sharp. Motivation in Software Engineering: A Systematic Literature Review. 2007.

David Benavides, Sergio Segura, and Antonio Ruiz-Cortés. Automated Analysis of Feature Models 20 Years Later: A Literature Review. *Information Systems*, 2010.

Lianping Chen, Muhammad Ali Babar, and Nour Ali. Variability Management in Software Product Lines: A Systematic Review. ACM, 2007.

T. Dyba, B. A. Kitchenham, and M. Jorgensen. Evidence-based software engineering for practitioners. *IEEE Software*, 22(1):58–65, Jan 2005.

Vigdis By Kampenes, Tore Dybå, Jo E. Hannay, and Dag I K Sjøberg. A systematic review of effect size in software engineering experiments. *Information and Software Technology*, 2007.

Khalid S Khan, Gerben Ter Riet, Julie Glanville, Amanda J Sowden, Jos Kleijnen, et al. *Undertaking systematic reviews of research on effectiveness: CRD's guidance for carrying out or commissioning reviews*. Number 4 (2n). NHS Centre for Reviews and Dissemination, 2001.

Barbara Kitchenham. Procedures for Performing Systematic Reviews. 2004.

Barbara Kitchenham and Pearl Brereton. A systematic review of systematic review process research in software engineering. *Information and Software Technology*, 55(12): 2049–2075, 2013.

Barbara Kitchenham and Stuart Charters. Guidelines for performing Systematic Literature Reviews in Software Engineering. Technical report, 2007.

Dag I K Sjøberg, Jo E. Hannay, Ove Hansen, Vigdis By Kampenes, Amela Karahasanović, Nils Kristian Liborg, and Anette C. Rekdal. A survey of controlled experiments in software engineering. *IEEE Transactions on Software Engineering*, 31(9):733–753, 2005.

Darja Šmite, Claes Wohlin, Tony Gorschek, and Robert Feldt. Empirical evidence in global software engineering: a systematic review. *Empir Software Eng*, 15:91–118, 2010.

Jane Webster and Richard T. Watson. Analysing the Past to Prepare for the Future: Writing a Literature Review. *MIS Quarterly*, 26(2), 2002.