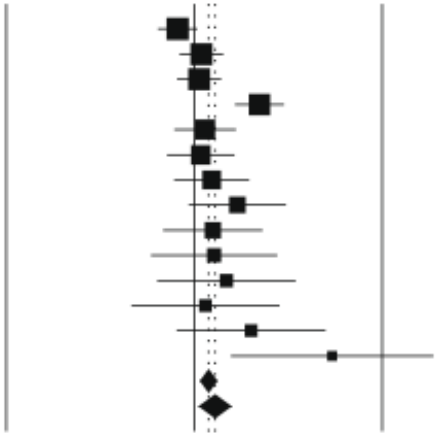
Durham University

"The idea is to try to give all the information to help others to judge the value of your contribution; not just the information that leads to judgement in one particular direction or another"
[Richard P Feynman]



Performing a Mapping Study

Background & Tutorial

David Budgen

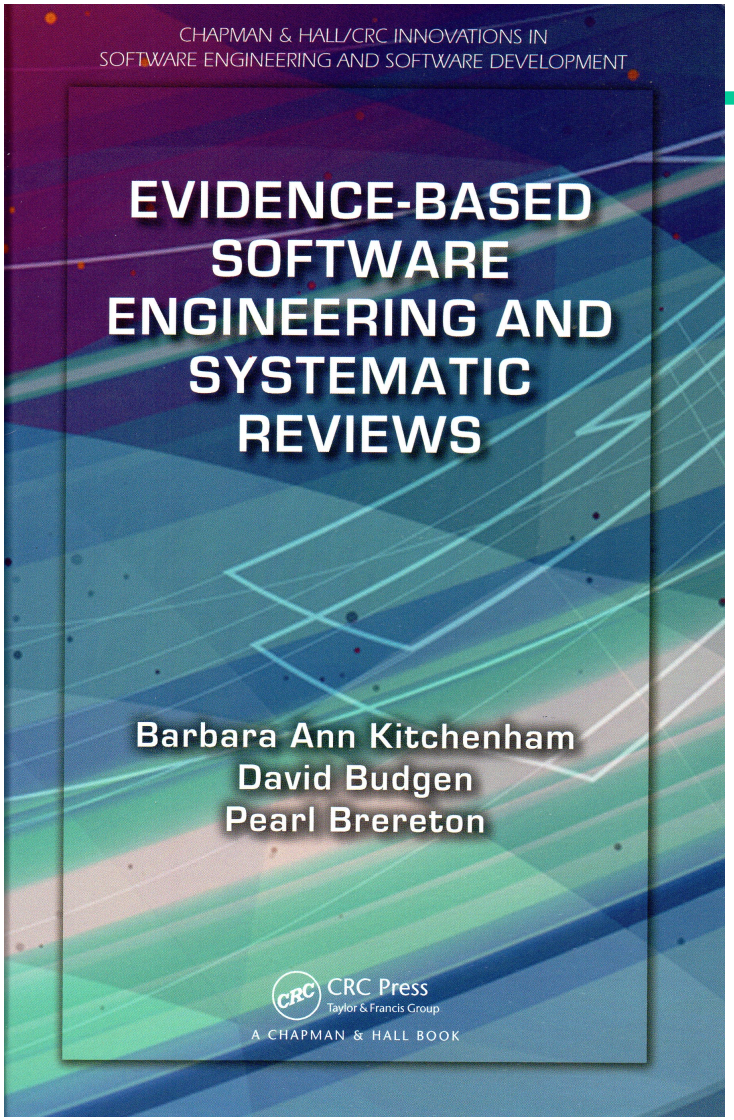
BACKGROUND

These slides were originally produced in support of a tutorial/ seminar given by David Budgen in Newcastle University, December 2022. They are provided on this site as a resource that might aid others who are planning to undertake a mapping study. If you reuse any of the information provided, please acknowledge the web site (or any of the original sources cited here).

Structure of my talk

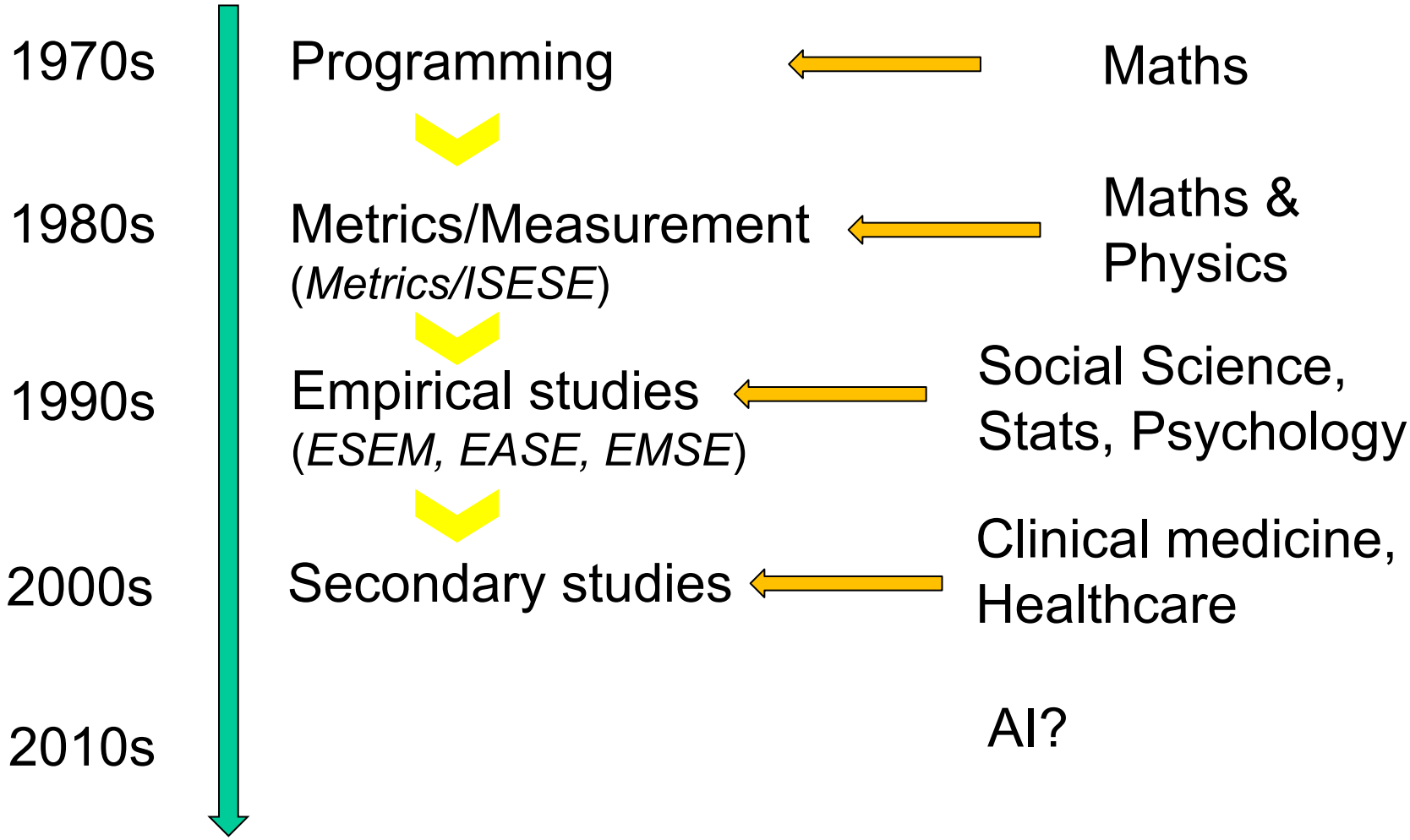
- Mapping Studies are a particular form of Secondary Study, and these are a particular form of *Empirical Study*, so I'll begin by briefly explaining a bit about empirical studies and their role(s).
- I'll then look at the form of a *Mapping Study*.
- Followed by discussion about each of the steps involved and some possible issues that can arise.

- I'll illustrate this with some examples from the software engineering literature.



This provides a lot of detailed information, but runs to around 400 pages, so I presume you would prefer a synopsis!

Empirical Studies: Evolution



Empirical Studies: Primary Studies

There are many forms of 'primary study' with varying degrees of 'controlled intervention' by the researcher.

- *Experiments* tend to be highly controlled
- *Observational* studies (such as ethnographical ones) may well be completely uncontrolled

The idea of a 'field study' is important (this is one that is performed in a real-life context).

- For clinical medicine this is usually an RCT (*Randomised Clinical Trial*)
- For software engineering it is more likely to be a *Case Study* (caveat: this term is not always used rigorously)

Primary or Secondary?

- In SE we conduct empirical evaluation through both primary and secondary forms of study:
  - in a *primary* study, we directly study the entity of interest (a technique, a way of structuring software, …) by making observations and measurements
  - in a *secondary* study, we seek to aggregate the outcomes of many different primary studies (to help overcome the inherent variability of individual studies)
- The vocabulary and forms of empirical study used in software engineering are largely adapted from other disciplines where such studies involve human participants: social science, psychology, education…

Secondary studies

- Essentially divide into two groups:
  - *Systematic Reviews* seek to find all the primary studies addressing a particular question, assess them for rigour and quality, and then aggregate their findings. When these are experiments, aggregation can employ statistical meta-analysis, but SE data is rarely good enough for this.
  - *Mapping Studies* seek to identify what studies exist that address a (generally broader) question, to categorise them and to identify 'gaps'. May precede a fuller systematic review for a topic if there are enough 'good' studies.
- Together, their use underpins what we term the *evidence-based* paradigm.

Quantitative studies

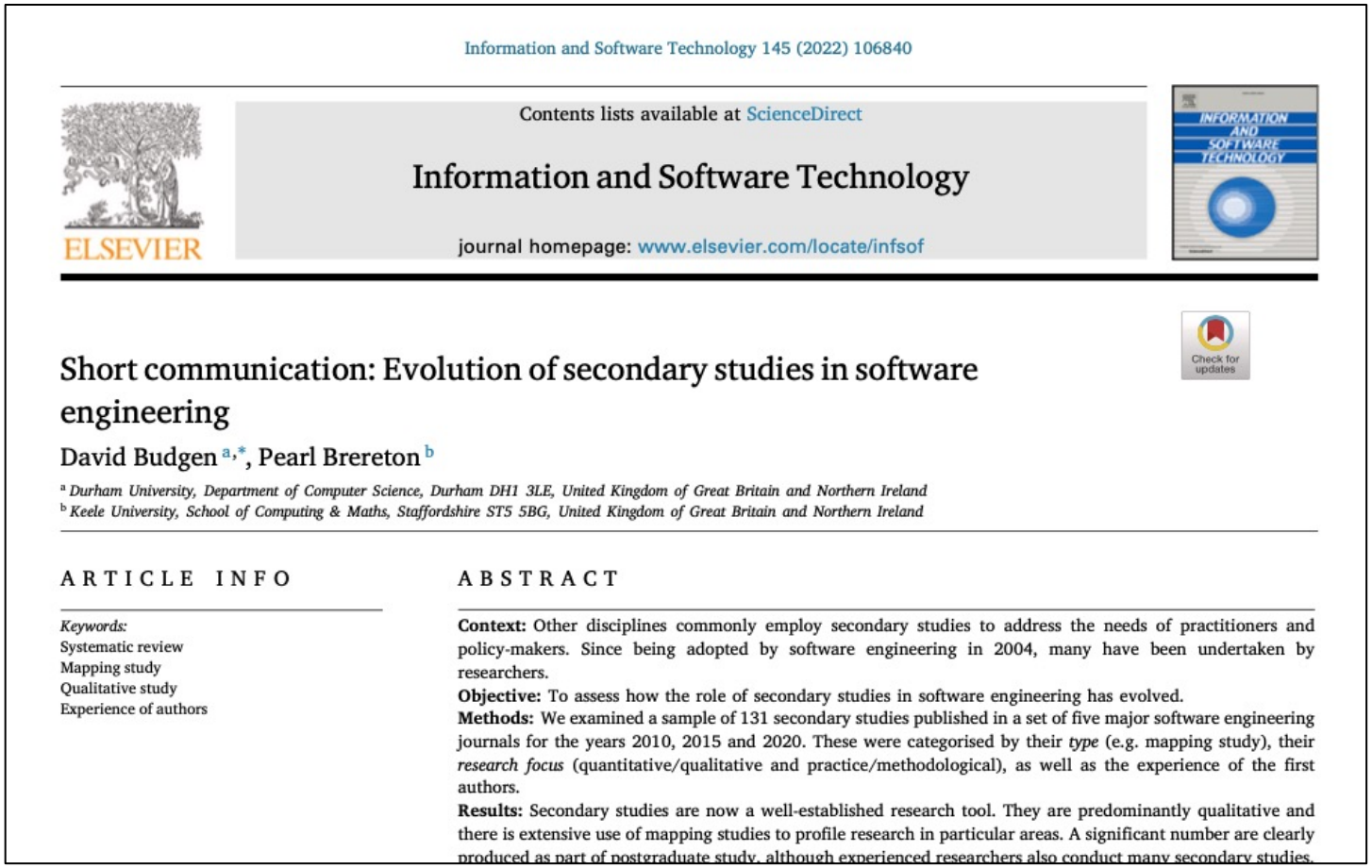
- *Quantitative* evaluation is widely used to determine whether a *cause-effect* relationship exists, and so:
  - may test the effect of some intervention (the *treatment*)
  - uses measures based on 'counting' scales (eg ratio scale)
  - may be able to employ statistical forms to aid analysis

- Example: "does using pair programming for complex tasks lead to faster development than when using solo programming?"

Qualitative studies

- *Qualitative* evaluation usually involves studying entities in their natural setting, often through some form of *observation*, hence:
  - analysis involves *interpretation* based on explanations
  - the process of analysis needs to recognise that there may be *different* interpretations

- Example: "why is it that different inspection groups may find more/fewer errors, and that they may also find different types of error?"



Information and Software Technology 145 (2022) 106840

Contents lists available at ScienceDirect

Information and Software Technology

journal homepage: www.elsevier.com/locate/infsof

Short communication: Evolution of secondary studies in software engineering

David Budgen [a,*], Pearl Brereton [b]

[a] Durham University, Department of Computer Science, Durham DH1 3LE, United Kingdom of Great Britain and Northern Ireland
[b] Keele University, School of Computing & Maths, Staffordshire ST5 5BG, United Kingdom of Great Britain and Northern Ireland

ARTICLE INFO

Keywords:
Systematic review
Mapping study
Qualitative study
Experience of authors

ABSTRACT

Context: Other disciplines commonly employ secondary studies to address the needs of practitioners and policy-makers. Since being adopted by software engineering in 2004, many have been undertaken by researchers.
Objective: To assess how the role of secondary studies in software engineering has evolved.
Methods: We examined a sample of 131 secondary studies published in a set of five major software engineering journals for the years 2010, 2015 and 2020. These were categorised by their *type* (e.g. mapping study), their *research focus* (quantitative/qualitative and practice/methodological), as well as the experience of the first authors.
Results: Secondary studies are now a well-established research tool. They are predominantly qualitative and there is extensive use of mapping studies to profile research in particular areas. A significant number are clearly produced as part of postgraduate study, although experienced researchers also conduct many secondary studies.

We looked at a sample of 131 secondary studies published in 2010, 2015, 2020. 77% were mapping studies, 16% were systematic reviews. 94% were qualitative studies of SE practice.

Variation in measurements

- Our data is also affected by various forms of *variation*.
- This variation is natural and unavoidable.
- We address the issue by drawing on the experiences of other disciplines such as clinical medicine, social science, psychology etc. These have developed rigorous empirical practices to minimize the bias that might be caused by this.

Why do measurements vary?

Natural Sciences (Humans as *Observers)*	Any variation in the results of experiments tends to come from errors in measurement and so are usually small and normally distributed.
Humans as *Recipients* of experimental treatment (Clinical RCTs)	We expect some 'spread' in the outcomes because humans differ physically, and also in the way that they respond to a treatment.
Humans as *Participants* (Software Engineering experiments and quasi-experiments, case studies etc.)	We expect a large 'spread' in the outcomes because each person involved will have different abilities, skills, and experience. Rather as we usually expect a class of students to receive a wide range of marks on a module.

Illustration largely relates to 'experiments', but the issue is generic. In the same way, software artifacts differ. Making measurements relating to people+software can therefore be challenging.

The Evidence-Based Paradigm

- The *evidence-based* paradigm originated in clinical medicine. A major stimulus came from the noted epidemiologist Archie Cochrane (1909-1988), who was concerned about the quality of the research evidence being used to inform practice and teaching.

- Evidence-Based Medicine (EBM) seeks to employ *secondary studies* as the tool for finding, judging and synthesizing the outcomes of **all** relevant empirical studies in order to draw conclusions about clinical treatments. It has had a major impact upon clinical practice and upon healthcare in general.

Evidence in medicine

- The logo of the *Cochrane Collaboration* illustrates the concept of pooling data, taken from a landmark study in New Zealand.
- The horizontal lines in the 'forest plot' represent the results from a series of 7 RCTs of an intervention used with pregnant women likely to give birth prematurely.
- Individually, only two of the studies showed some benefits from the treatment
- The diamond at the bottom shows the result of a meta-analysis conducted 8 years later, strongly indicating clear benefit from the intervention – and as Ben Goldacre observes "we should always remember the human cost of these numbers"

In a forest plot, each primary study is represented by a horizontal line, with the width indicating the variance in its results. If the end of the horizontal line is to the left of the vertical line, it means the treatment was better than a placebo, but if the horizontal line touches the vertical line, then the results show no clear difference (statistically).

What is *evidence*?

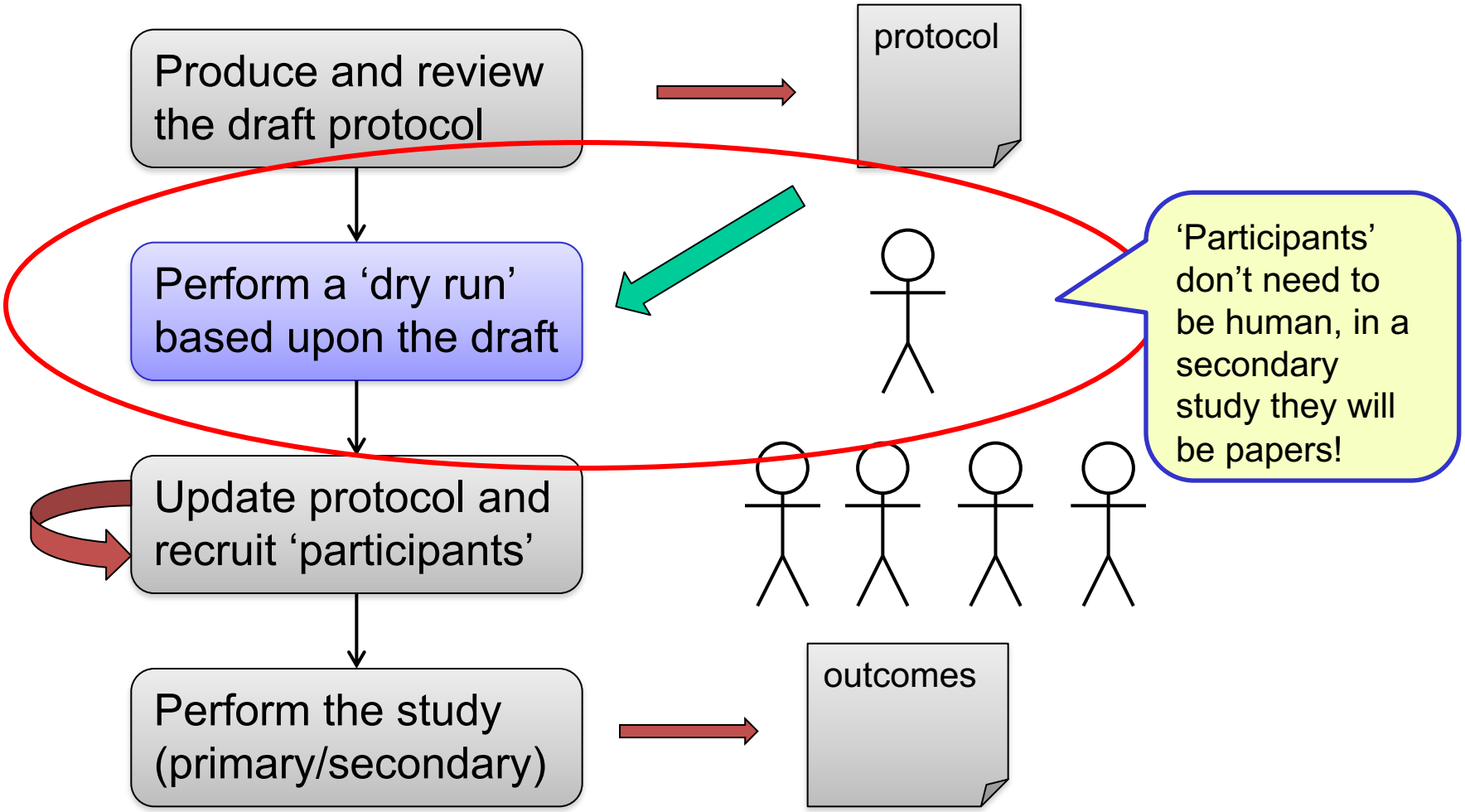
- We consider *evidence* to be derived from the aggregated outcomes of many *primary* studies
  - reinforcing findings that are common
  - reducing the effect of variability/bias in individual studies.
- In particular, the process of identifying well-founded evidence by using a *secondary study* requires:
  - comprehensive and exhaustive searches to find *all* potentially relevant primary studies
  - using carefully defined procedures for deciding whether to include or *exclude* each study that is found
- Aim is to minimise bias and to emphasise the objectivity of the procedures employed.

Now to some practicalities

The research protocol

- Any empirical evaluation *process* needs to be:
  - *objective*
  - *unbiased*
- And analysis should avoid 'fishing' for results from the outcomes. So, we begin by creating a *plan* for conducting the study, termed the research protocol.
  - Usually perform some form of 'dry run' of study elements to test the protocol in a controlled situation.
  - When reporting the study, we also need to describe any *divergences* from the plan that occurred (and why).
  - The protocol may also identify likely *risks of bias*: factors that we can't control and that might reduce our confidence in the outcomes.

Role of the *dry run*



Planning secondary studies

- For a *secondary study* the research protocol should:
  - ❑ specify a well-focused *research question*
  - ❑ use the RQ to identify a set of *keywords* for searching
  - ❑ specify how and where to *search* for source material (may be manual and/or electronic); and the period to search
  - ❑ provide clear *inclusion/exclusion* rules for selecting primary studies
  - ❑ identify a suitable means for performing *aggregation*
- A *mapping study* differs from a systematic review mainly in the broader nature of the research question addressed, and in performing *categorisation* rather than aggregation.

Benefits of a protocol

- Good planning avoids wasting time or having to repeat parts of the study. ✅
- A dry run should involve performing all of the activities to some degree if possible -- including analysis. (See next slide!)

- AND, when you come to write a paper, you can reuse lots of the text from the protocol when describing your empirical method. (Just remember to edit the tenses…) ✅✅✅

ELSEVIER

The Journal of Systems and Software 67 (2003) 55–75

The Journal of Systems and Software

www.elsevier.com/locate/jss

CASE tool evaluation: experiences from an empirical study

David Budgen *, Mitchell Thomson [1]

Department of Computer Science, Keele University, Staffordshire ST5 5BG, UK

Received 1 July 2001; received in revised form 10 September 2001; accepted 21 February 2002

Abstract

While research activity in software engineering often results in the development of software tools and solutions that are intended to demonstrate the feasibility of an idea or concept, any resulting conclusions about the degree of success attained are rarely substantiated through the use of supporting experimental evidence. As part of the development of a prototype computer assisted software engineering (CASE) tool intended to support opportunistic software design practices, we sought to evaluate the use of the tool by both experienced and inexperienced software engineers. This work involved performing a review of suitable techniques, and then designing and performing a set of experimental studies to obtain data which could be used to assess how well the CASE tool met its design goals. We provide an assessment of how effective the chosen evaluation process was, and conclude by identifying the need for an 'evaluation framework' to help with guiding such studies.

This *could* have been a much more insightful study. However, we failed to perform the analysis in our dry run – missing a problem with data logging that could have been easily rectified. This resulted in our eventual analysis being much weaker than it might have been.

Procedures for a Secondary study

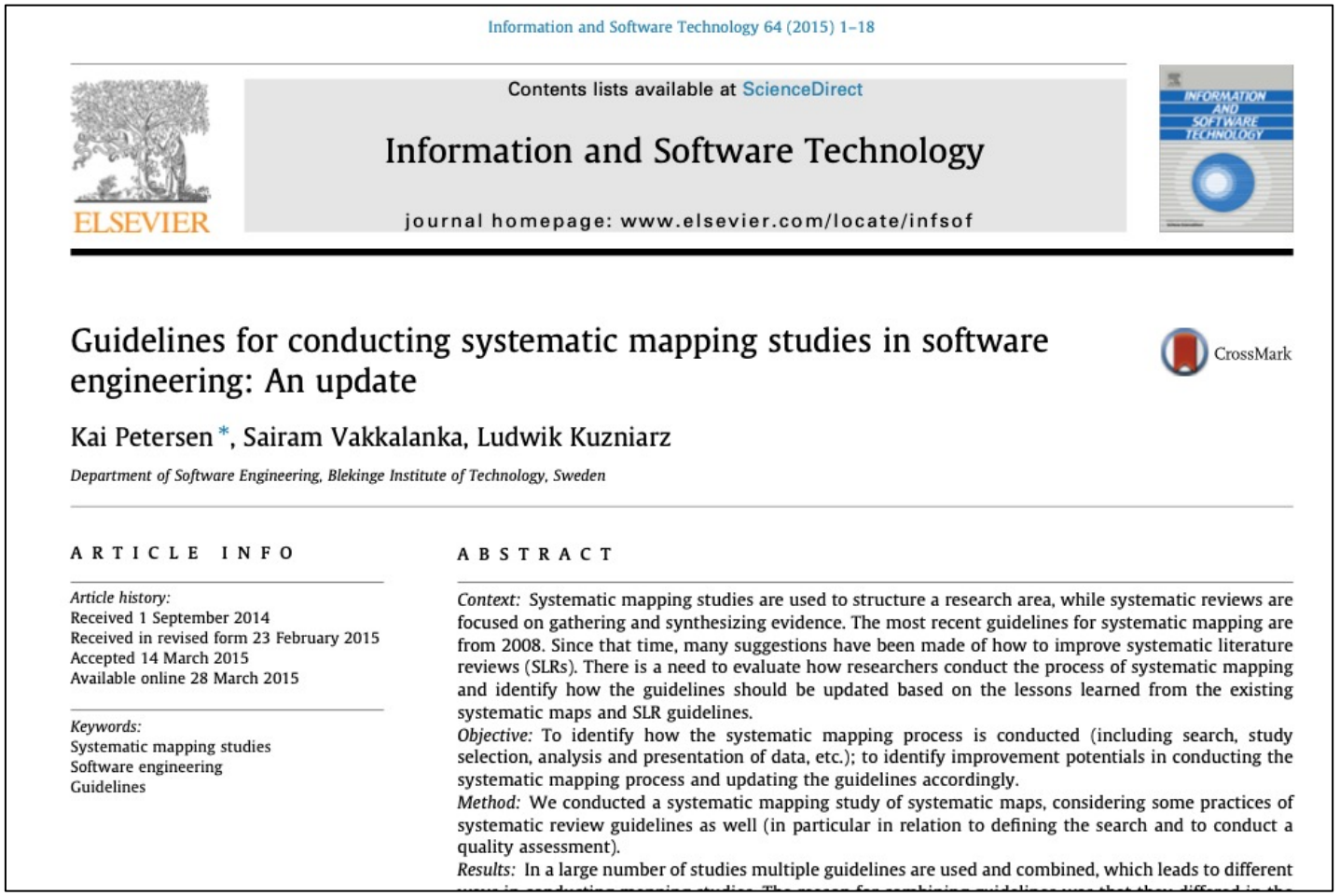
Phase 1: Plan review

- specify research question used to create search strings
- develop review protocol (plan)
- validate protocol, which may include prototyping search strings

Phase 2: Conduct review

- execute search strategy: strings, sources, bounding dates etc.
- select primary studies: title, abstract, full paper
- assess study quality (often omitted for mapping studies)
- perform data extraction
- synthesise the data to answer the research question

Phase 3: Document the outcomes



Information and Software Technology 64 (2015) 1–18

Contents lists available at ScienceDirect

Information and Software Technology

journal homepage: www.elsevier.com/locate/infsof

Guidelines for conducting systematic mapping studies in software engineering: An update

Kai Petersen *, Sairam Vakkalanka, Ludwik Kuzniarz

Department of Software Engineering, Blekinge Institute of Technology, Sweden

ARTICLE INFO

Article history:
Received 1 September 2014
Received in revised form 23 February 2015
Accepted 14 March 2015
Available online 28 March 2015

Keywords:
Systematic mapping studies
Software engineering
Guidelines

ABSTRACT

Context: Systematic mapping studies are used to structure a research area, while systematic reviews are focused on gathering and synthesizing evidence. The most recent guidelines for systematic mapping are from 2008. Since that time, many suggestions have been made of how to improve systematic literature reviews (SLRs). There is a need to evaluate how researchers conduct the process of systematic mapping and identify how the guidelines should be updated based on the lessons learned from the existing systematic maps and SLR guidelines.

Objective: To identify how the systematic mapping process is conducted (including search, study selection, analysis and presentation of data, etc.); to identify improvement potentials in conducting the systematic mapping process and updating the guidelines accordingly.

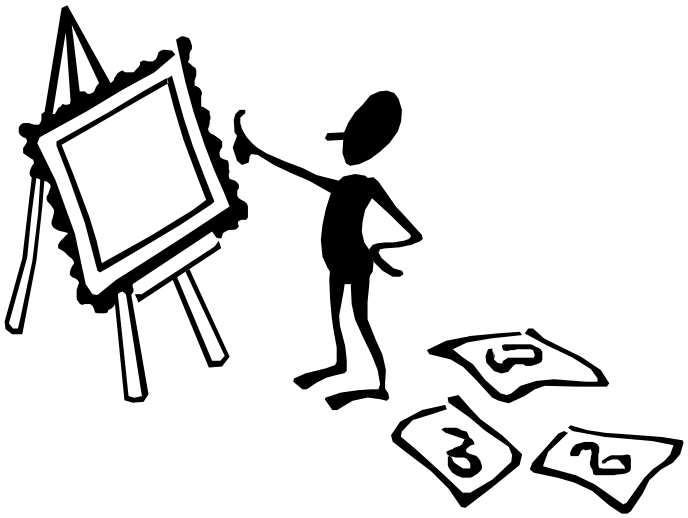
Method: We conducted a systematic mapping study of systematic maps, considering some practices of systematic review guidelines as well (in particular in relation to defining the search and to conduct a quality assessment).

Results: In a large number of studies multiple guidelines are used and combined, which leads to different

A useful source of guidance about performing mapping studies.

Phase 1: Plan Review

- Need to carefully formulate the *research questions* as these are used to construct the search strings used to search literature databases
- Experience suggests that these may go through several stages of refinement
- As with primary studies, may need to *pilot* key elements such as search strings through some form of ‘dry run’
- Can then write a *review protocol* that will guide the process of review



Example research questions

We ask the following research question (RQ) and the corresponding sub-questions:

RQ: How has empirical research on human aspects of software architecture decision making been done so far and what can we learn from that?

We detail this question into three aspects: the focus of the studies, their objective and the research design. The first two allow us to position the studies according to their insights about software architecture:

RQ1.1: What are the study foci and how to characterize them?
RQ1.2: What are the objectives of the studies and how do they relate to the focus?
RQ1.3: What is the research design of the studies?

Example of research question and expansion into sub-questions. (Razavian et al., JSS 149, 2019)

Experiences (SE): Planning

- We usually use version numbering and have a ‘change table’ at the start of the protocol document to indicate how it has evolved.
- Good idea is to get an independent reviewer to look at it, perhaps as a walk-through.

- Key issues to address:
  - ❑ Are search strings appropriately derived from RQs?
  - ❑ Will the data extracted properly address the RQs?
  - ❑ Does the data analysis procedure answer the RQs?
  - ❑ Do procedures as planned provide adequate rigour?

Example protocol header

Protocol for a Study of the Experience of First Authors of Secondary Studies for Software Engineering Topics

David Budgen

This protocol forms a plan for analysing the profiles of the leading authors of software engineering systematic reviews and like studies, using a sample of published secondary studies from three well-spaced years, drawn from five journals.

Change Table

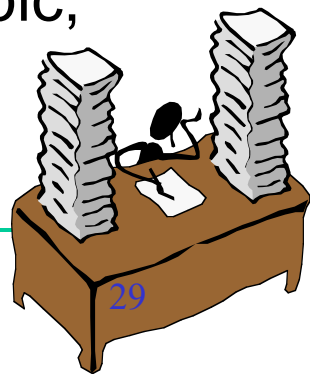
Version	*Date*	*Changes*	*Reasons for Change*
1.0	24 February 2021	First draft	
1.1	25 February 2021	Added 2010; removed post-publication counts.	2010 was relatively early in the time when systematic reviews became used and although the sample is small, it appears to involve more established researchers.
1.2	10 March 2021	Refined counts; retitled study as being of 'secondary studies'	BAK suggested extending counts; papers found indicated title was too narrow
1.3	15 March 2021	Section 1 & add RQ, more on analysis and data collection.	Clarify likely effects of having inexperienced researchers conduct systematic reviews.
1.4	25 March 2021	Revise counting rules and reorganise sections 5 and 9 for consistency.	Trial of protocol using DBLP and the authors from 2010 conducted by both members of the team.

This protocol is relatively short (four pages) but still went through several iterations

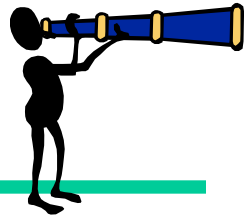
Version 1.4 was motivated by performing a dry run.

Phase 2: Identification of studies

- This involves executing the search strategy defined in the review protocol using multiple search engines.
- Some issues in this:
  - *keywords* to use (and combinations of these)
  - where to search, journals and conferences in SE are major sources, but might also want to search the unpublished *grey literature* (technical reports, non-refereed material etc.)
  - what *dates* to use in bounding the search (helps if some paper or book can be considered as forming a baseline)
- Outcome may well be that we identify a large number of documents, actual number depends on the topic, but but probably hundreds/few thousand.



Experiences (SE): searching



- Our own experiences of the available search engines (such as IEEExplore, ACM, Web of Science, Google Scholar,…) is that:
  - they each tend to search a different subset of sources and there may be little overlap in what they find
  - The organisation of the boolean combinations of the keywords is different for each one, and they do not always generate consistent results
  - even collectively, they do not find all of the papers
- We mainly use Google Scholar for prototyping possible terms (has poor filters for more serious work)

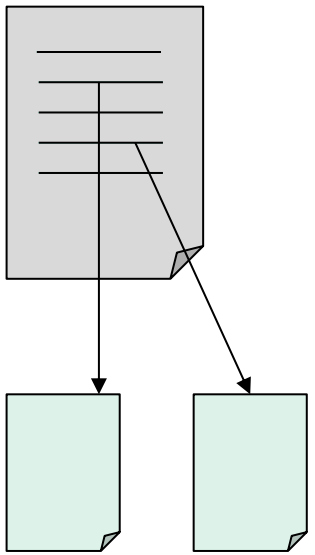
Completeness

- For a systematic review the results from a search needs to be as complete as possible.
- For mapping studies the aim is to get a ‘good sample’, but of course we don’t know the size of the complete set!

- We usually recommend having a *gold standard* to help assess effectiveness of search. This consists of a small number of known papers that the searching process should identify. Ideally these will be by different authors.

Snowballing

- Involves taking the set of selected papers after inclusion/exclusion and working through the references in these to see if they identify relevant papers that our search has missed. (*Backwards snowballing*.)
- Can be regarded as a form of secondary search.
- Anecdotally might expect to increase the size of the selected set by as much as an additional 10%.



Manual search…

- …is a useful option in some cases.

- Our study of authors used manual search for secondary studies. This was revealing!
  - Authors don't always mention the topic or type of study in title/abstract
  - Some secondary studies described a mapping study in the 'method' section but never used any of the EBSE/systematic review terminology or references

- We did back it up with an electronic search.
- Hint: can also be used to create a 'gold standard'.

Removing duplicates

- When using multiple search engines can expect the same papers to be found in more than one of them. So, need to remove these *duplicates*.

- Another source of duplicates is the common situation where the authors publish a conference paper and then a (more detailed) journal paper. However, this should only count as one study.

- Can be quite hard to spot. [Hint: look at the counts for participants/documents etc.] Authors may also write more than one conference paper using same data!

Phase 2: Select & Assess

- *Selection* of relevant papers (*inclusion/exclusion*) involves sifting through the list of candidate papers, examining:
  - the *title*, and if that looks promising…
  - the *abstract*, and if that continues to look promising…
  - the *actual* paper
- Ideally, this is done independently by two researchers and the protocol should identify the procedure(s) to follow if they disagree.
- *Assessment* of study quality is largely concerned with how the primary study being reported was conducted, and how well this is reported (not mapping studies)

Select & Assess: First filter (title)

- If you have a lot of papers from the search, start with just the title.
- It may be that just one person can do this, removing any papers that are obviously irrelevant. If in any doubt, leave a paper in for a fuller examination. However, better if two do it and compare results.

- [Personal view, this might be harder for mapping studies than for systematic reviews.]
- *NB Keep records of all decisions throughout*. Can use to calculate level of agreement (*Kappa test*) between analysts.

Select & Assess: Next filters

- These really need two people working independently.
- Reading abstracts reveals how uninformative many of these are!
- In empirical SE we encourage the use of structured abstracts (*context, objective, method, results, conclusion*) so that essential information is available.
- If it is necessary to read the paper to decide whether to include it, often only need to consult the introduction section, possibly the conclusions.

Selection: working on your own

- Where there is only one person to undertake these tasks (not ideal, but…), two options that can be employed to improve rigour are:
  - one person decides/extracts, while a second person (e.g. PhD supervisor) checks
  - use *test-retest*: this involves assessing all the studies, and then re-assessing them after a suitable time interval

- If you use test-retest, need to calculate degree of agreement (use a *Kappa test*).

Inclusion/Exclusion: example

- The study should be written in English.
- The study should be published between 2006 and December 2016.
- The study directly answers one or more of the research questions of this study.
- The study should clearly state its focus on Kanban in the software engineering domain.
- The study should describe the elements and the approach used to implement Kanban.
- If the study has been published in more than one journal or conference, the most recent version of the study is included.

Studies were excluded if their focus was not specifically Kanban or if they did not provide academic rigour or industry relevance. The exclusion criteria used was:

- Short papers.
- Duplicate articles.
- Not written in English.
- Simulation studies.
- Studies not clearly focused on Kanban in the software engineering domain (e.g. industrial engineering, manufacturing and automotive industry).
- Not peer-reviewed scientific papers (i.e. books, book chapters, articles).

Taken from Ahmad et al, ”Kanban in Software Engineering, A Systematic Mapping Study”, JSS, 137, 2018.

Grey Literature

872 IEEE TRANSACTIONS ON SOFTWARE ENGINEERING, VOL. 49, NO. 2, FEBRUARY 2023

How Should Software Engineering Secondary Studies Include Grey Material?

Barbara Kitchenham, *Member, IEEE*, Lech Madeyski, *Senior Member, IEEE*, and David Budgen, *Member, IEEE*

Abstract—*Context*: Recent papers have proposed the use of *grey literature* (GL) and multivocal reviews. These papers have raised issues about the practices used for systematic reviews (SRs) in software engineering (SE) and suggested that there should be changes to the current SR guidelines. *Objective*: To investigate whether current SR guidelines need to be changed to support GL and multivocal reviews. *Method*: We discuss the definitions of GL and the importance of GL and of industry-based field studies in SE SRs. We identify properties of SRs that constrain the material used in SRs: a) the nature of primary studies; b) the requirements of SRs to be auditable, traceable, and reproducible; and explain why these requirements restrict the use of blogs in SRs. *Results*: SR guidelines have always considered GL as a possible source of primary studies and have never supported exclusion of field studies that incorporate the practitioners' viewpoint. However, the concept of GL, which was meant to refer to documents that were not formally published, is now being extended to information from sources such as blogs/tweets/Q&A posts. Thus, it might seem that SRs do not make full use of GL because they do not include such information. However, the unit of analysis for an SR is the primary study. Thus, it is not the *source* but the *type* of information that is important. Any report describing a rigorous empirical evaluation is a candidate primary study. Whether it is actually included in an SR depends on the SR eligibility criteria. However, any study that cannot be guaranteed to be publicly available in the long term should not be used as a primary study in an SR. This does not prevent such information from being aggregated in surveys of social media and used in the context of evidence-based software engineering (EBSE). *Conclusions*: Current guidelines for SRs do not require extensions, but their scope needs to be better defined. SE researchers require guidelines for analysing social media posts (e.g., blogs, tweets, vlogs), but these should be based on qualitative primary (not secondary) study guidelines. SE researchers can use mixed-methods SRs and/or the fourth step of EBSE to incorporate findings from social media surveys with those from SRs and to develop industry-relevant recommendations.

Our paper argues that only unpublished formal studies should be included, but for mapping studies might be a bit broader?

Example selection process

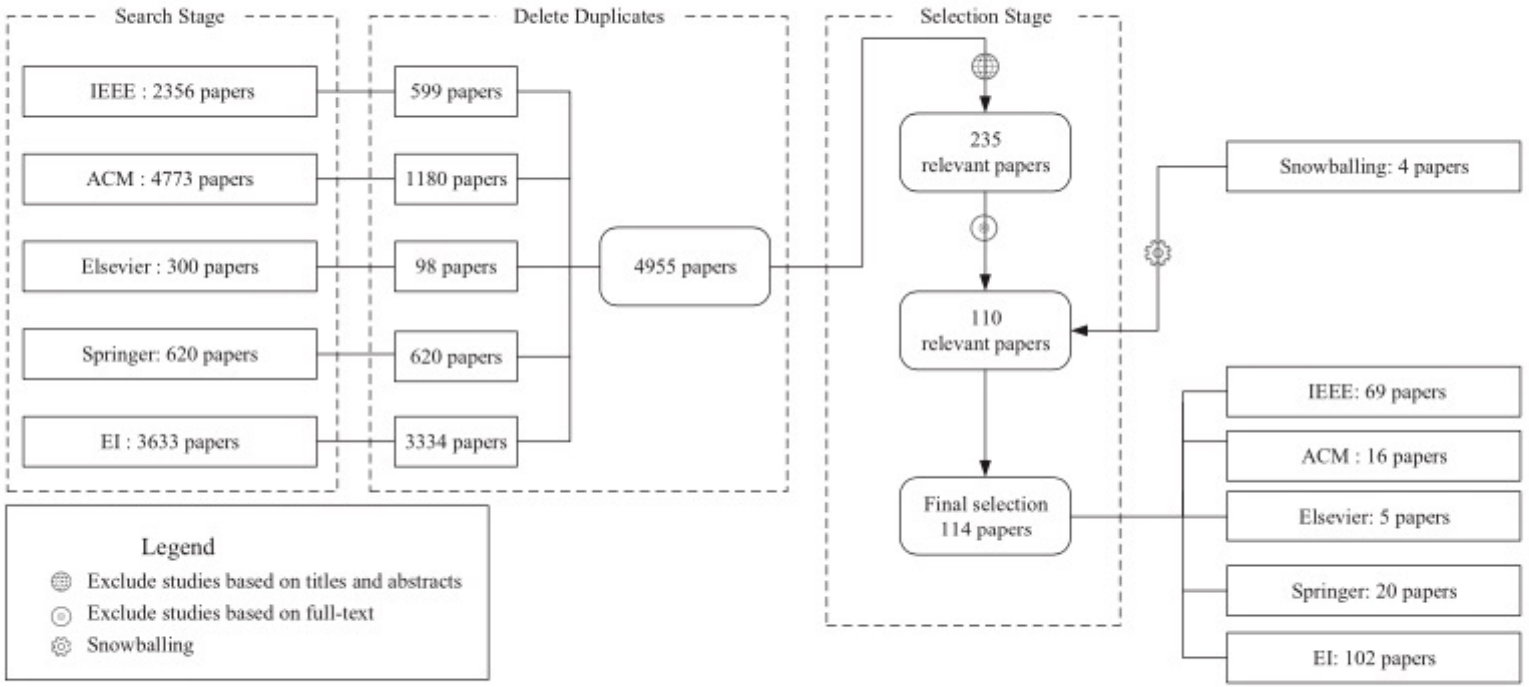


Fig. 2. Study search and selection results.

Phase 2: Data Extraction

- One of the tasks when writing a protocol is the design of the *data extraction forms* used to record the outcomes from a paper.
- Aim is to make data extraction as easy as possible and also to remind the analyst what is to be noted.

- Again, best done by two people, comparing their findings (and performing a *Kappa test* to determine the level of agreement)

Extraction: experience

- Don't expect authors to make this easy for you. Because there is no standard format for papers, relevant information may appear anywhere in a paper, not just in a *Results* section.

- *Personal illustration* – finding a key item of information related to a study in the caption of one of the figures (it was a count). It didn't appear anywhere else in the paper.

Categorisation

- Here's where mapping studies differ from systematic reviews. Also, need to categorise what is *relevant* to the RQ.

- Some early mapping studies were very weak in terms of this. Do we really need to know which countries produced most of the papers; where they were published; who were the most prolific authors etc.?

- *BUT* we might want to know if a particular research group produced most of the studies in a particular category, since this could be a source of bias.

Categorisation: experience

- Can be quite hard to perform – a paper might report several different aspects from a range of categories and possibly provide results from multiple studies.
- Make sure you have an 'other' category wherever appropriate.
- *Topic-independent* classification can use existing models, but even then, for classification of (say) research methods there may be different models.
- *Topic-dependent* classification can be based on domain models or can emerge from the study.

Categorisation: visualizing results

- Text mining and content analysis tools can help to identify clusters of studies. Felizardo has explored visualization techniques in a number of papers.
- Bubble plots may be useful (example below).

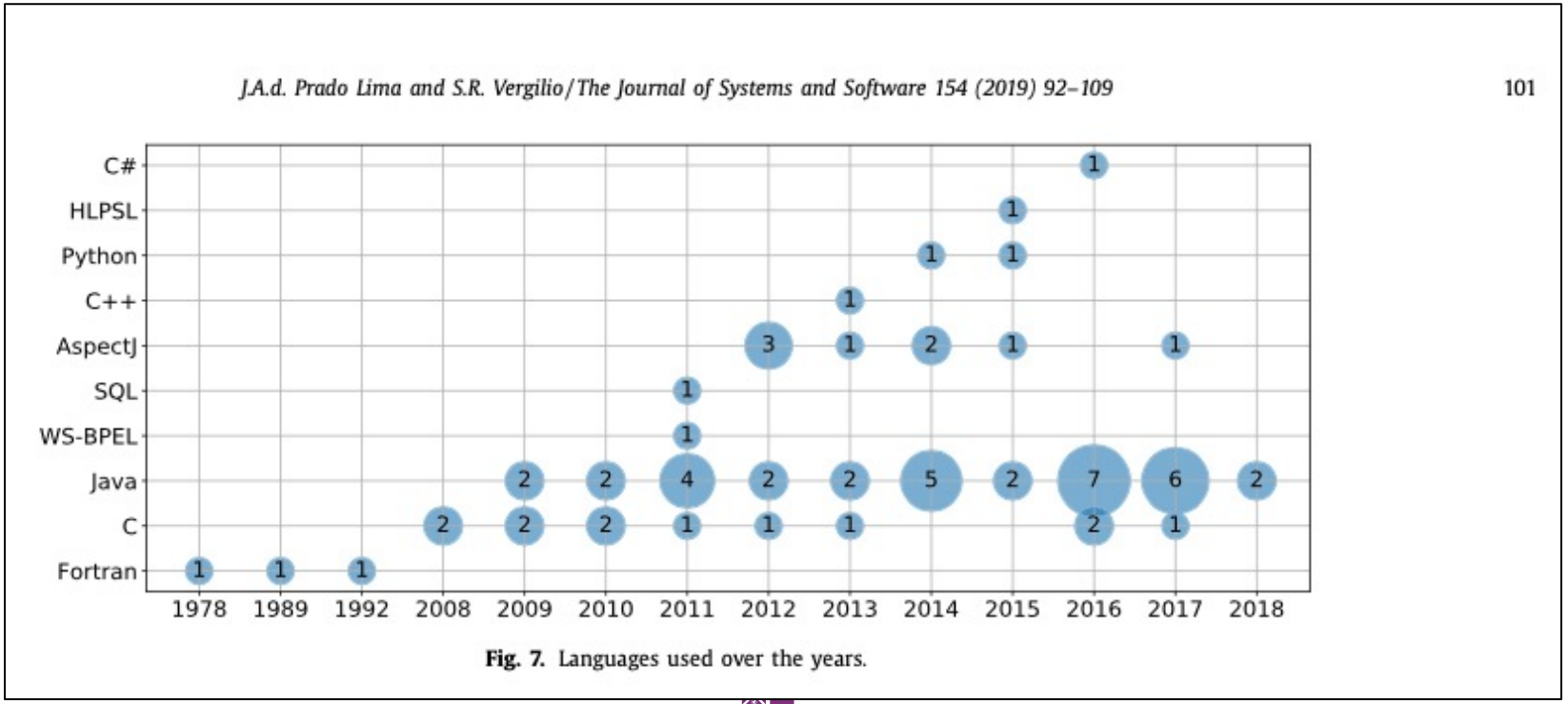


Fig. 7. Languages used over the years.

Experience: reporting

- Reporting standards in SE are generally poor. In particular, when extracting data from primary studies:
  - *abstracts* are often poorly written and omit information needed to determine whether or not a paper is relevant
  - papers tend to report only the data relevant to the immediate research question being addressed in the paper (if that), and few authors have any sense of adding something to an existing corpus of knowledge by including all relevant information
- To some extent this is probably because computing has only recently embraced the use of secondary studies -- so authors don't think about possible users when writing abstracts for primary studies.

Reporting: SEGRESS Guidelines

IEEE TRANSACTIONS ON SOFTWARE ENGINEERING, VOL. 49, NO. 3, MARCH 2023

SEGRESS: Software Engineering Guidelines for REporting Secondary Studies

Barbara Kitchenham, *Member, IEEE*,
Lech Madeyski, *Senior Member, IEEE*, and David Budgen, *Member, IEEE*

Abstract—*Context*: Several tertiary studies have criticized the reporting of software engineering secondary studies. *Objective*: Our objective is to identify guidelines for reporting software engineering (SE) secondary studies which would address problems observed in the reporting of software engineering systematic reviews (SRs). *Method*: We review the criticisms of SE secondary studies and identify the major areas of concern. We assess the PRISMA 2020 (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) statement as a possible solution to the need for SR reporting guidelines, based on its status as the reporting guideline recommended by the Cochrane Collaboration whose SR guidelines were a major input to the guidelines developed for SE. We report its advantages and limitations in the context of SE secondary studies. We also assess reporting guidelines for mapping studies and qualitative reviews, and compare their structure and content with that of PRISMA 2020. *Results*: Previous tertiary studies confirm that reports of secondary studies are of variable quality. However, *ad hoc* recommendations that amend reporting standards may result in unnecessary duplication of text. We confirm that the PRISMA 2020 statement addresses SE reporting problems, but is mainly oriented to quantitative reviews, mixed-methods reviews and meta-analyses. However, we show that the PRISMA 2020 item definitions can be extended to cover the information needed to report mapping studies and qualitative reviews. *Conclusions*: In this paper and its Supplementary Material, we present and illustrate an integrated set of guidelines called SEGRESS (Software Engineering Guidelines for REporting Secondary Studies), suitable for quantitative systematic reviews (building upon PRISMA 2020), mapping studies (PRISMA-ScR), and qualitative reviews (ENTREQ and RAMESES), that addresses reporting problems found in current SE SRs.

Index Terms—Evidence-based software engineering, reporting guidelines, systematic reviews, quality reviews, mapping studies, mixed-methods reviews, threats to validity, risk of bias, quality assessment, PRISMA 2020

Risk of Bias (RoB)

- Empirical studies usually assess *threats to validity* that might arise from the way that the study was conducted, or from external factors beyond control.

- For secondary studies, we now recommend adopting the term *Risk of Bias*, as used in other disciplines, and being more indicative of the ways in which the findings can be affected.

Some key points

- Secondary studies are an important tool for analysis of multiple studies (not necessarily human-centric).
- A sound and thorough *research protocol* is essential in order to ensure appropriate rigour.
- Performing a secondary study does need to be thorough and disciplined, and carefully documented.
- But, don't expect the authors of primary studies to be disciplined in terminology, organization of papers, or completeness of information!

- Enjoy…