Guidelines for performing Systematic Literature Reviews in Software Engineering

Barbara Kitchenham, und Stuart Charters. EBSE 2007-001. Keele University and Durham University Joint Report, (2007)

Source: „Guidelines for performing Systematic Literature Reviews in SE“, Kitchenham et al., 2007

Executive summary

The objective of this report is to propose comprehensive guidelines for systematic literature reviews appropriate for software engineering researchers, including PhD students. A systematic literature review is a means of evaluating and interpreting all available research relevant to a particular research question, topic area, or phenomenon of interest. Systematic reviews aim to present a fair evaluation of a research topic by using a trustworthy, rigorous, and auditable methodology.

The guidelines presented in this report were derived from three existing guidelines used by medical researchers, two books produced by researchers with social science backgrounds and discussions with researchers from other disciplines who are involved in evidence-based practice. The guidelines have been adapted to reflect the specific problems of software engineering research.

The guidelines cover three phases of a systematic literature review: planning the review, conducting the review and reporting the review. They provide a relatively high level description. They do not consider the impact of the research questions on the review procedures, nor do they specify in detail the mechanisms needed to perform meta-analysis.

Source: „Guidelines for performing Systematic Literature Reviews in SE“, Kitchenham et al., 2007

Examples of SLRs

R.F. Barcelos, G.H. Travassos, Evaluation approaches for software architectural documents: a systematic review, in: Ibero-American Workshop on Requirements Engineering and Software Environments (IDEAS), La Plata, Argentina, 2006.

T. Dyba, V.B. Kampenes, D.I.K. Sjøberg, A systematic review of statistical power in software engineering experiments, Information and Software Technology 48 (8) (2006) 745–755.

D. Galin, M. Avrahami, Do SQA programs work – CMM works. A meta analysis, IEEE International Conference on Software – Science, Technology and Engineering (2005).

D. Galin, M. Avrahami, Are CMM program investments beneficial? Analyzing past studies, IEEE Software 23 (6) (2006) 81–87.

J.E. Hannay, D.I.K. Sjøberg, T. Dybå, A systematic review of theory use in software engineering experiments, IEEE Transactions on SE 33 (2) (2007) 87– 107.

M. Jørgensen, Estimation of software development work effort: evidence on expert judgement and formal models, International Journal of Forecasting 3 (3) (2007) 449–462.

N. Juristo, A.M. Moreno, S. Vegas, Reviewing 25 years of testing technique experiments, Empirical Software Engineering Journal (1–2) (2004) 7–44.

Source: „Guidelines for performing Systematic Literature Reviews in SE“, Kitchenham et al., 2007

Glossary 1/2

Meta-analysis. A form of secondary study where research synthesis is based on quantitative statistical methods.

Primary study. (In the context of evidence) An empirical study investigating a specific research question.

Secondary study. A study that reviews all the primary studies relating to a specific research question with the aim of integrating/synthesising evidence related to a specific research question.

Sensitivity analysis. An analysis procedure aimed at assessing whether the results of a systematic literature review or a meta-analysis are unduly influenced by a small number of studies. Sensitivity analysis methods involve assessing the impact of high leverage studies (e.g. large studies or studies with atypical results), and ensuring that overall results of a systematic literature remain the same if low quality studies (or high quality) studies are omitted from the analysis, or analysed separately.

Systematic literature review (also referred to as a systematic review). A form of secondary study that uses a well-defined methodology to identify, analyse and interpret all available evidence related to a specific research question in a way that is unbiased and (to a degree) repeatable.

Source: „Guidelines for performing Systematic Literature Reviews in SE“, Kitchenham et al., 2007

Glossary 2/2

Systematic review protocol. A plan that describes the conduct of a proposed systematic literature review.

Systematic mapping study (also referred to as a scoping study). A broad review of primary studies in a specific topic area that aims to identify what evidence is available on the topic.

Tertiary study (also called a tertiary review). A review of secondary studies related to the same research question.

Source: „Guidelines for performing Systematic Literature Reviews in SE“, Kitchenham et al., 2007

Reasons for performing a SLR

There are many reasons for undertaking a systematic literature review. The most common reasons are:

- To summarise the existing evidence concerning a treatment or technology e.g. to summarise the empirical evidence of the benefits and limitations of a specific agile method.
- To identify any gaps in current research in order to suggest areas for further investigation.
- To provide a framework/background in order to appropriately position new research activities.

Source: „Guidelines for performing Systematic Literature Reviews in SE“, Kitchenham et al., 2007

The importance of SLRs

Most research starts with a literature review of some sort. However, unless a literature review is thorough and fair, it is of little scientific value. This is the main rationale for undertaking systematic reviews. A systematic review synthesises existing work in a manner that is fair and seen to be fair. For example, systematic reviews must be undertaken in accordance with a predefined search strategy. The search strategy must allow the completeness of the search to be assessed. In particular, researchers performing a systematic review must make every effort to identify and report research that does not support their preferred research hypothesis as well as identifying and reporting research that supports it.

Source: „Guidelines for performing Systematic Literature Reviews in SE“, Kitchenham et al., 2007

Advantage and disadvantages of SLRs

The advantages of systematic literature reviews are that:

- The well-defined methodology makes it less likely that the results of the literature are biased, although it does not protect against publication bias in the primary studies.
- They can provide information about the effects of some phenomenon across a wide range of settings and empirical methods. If studies give consistent results, systematic reviews provide evidence that the phenomenon is robust and transferable. If the studies give inconsistent results, sources of variation can be studied.
- In the case of quantitative studies, it is possible to combine data using meta-analytic techniques. This increases the likelihood of detecting real effects that individual smaller studies are unable to detect.

The major disadvantage of systematic literature reviews is that they require considerably more effort than traditional literature reviews. In addition, increased power for meta-analysis can also be a disadvantage, since it is possible to detect small biases as well as true effects.

Source: „Guidelines for performing Systematic Literature Reviews in SE“, Kitchenham et al., 2007

Features of SLRs

Some of the features that differentiate a systematic review from a conventional expert literature review are:

- Systematic reviews start by defining a review protocol that specifies the research question being addressed and the methods that will be used to perform the review.
- Systematic reviews are based on a defined search strategy that aims to detect as much of the relevant literature as possible.
- Systematic reviews document their search strategy so that readers can assess their rigour and the completeness and repeatability of the process (bearing in mind that searches of digital libraries are almost impossible to replicate).
- Systematic reviews require explicit inclusion and exclusion criteria to assess each potential primary study.
- Systematic reviews specify the information to be obtained from each primary study including quality criteria by which to evaluate each primary study.
- A systematic review is a prerequisite for quantitative meta-analysis.

Source: „Guidelines for performing Systematic Literature Reviews in SE“, Kitchenham et al., 2007

Comparing Software Engineering experimental methodology with that of other disciplines

Discipline	**Comparison with SE (1 is perfect agreement, 0 is complete disagreement)**
Nursing & Midwifery	0.83
Primary Care	0.33
Organic Chemistry	0.83
Empirical Psychology	0.66
Clinical Medicine	0.17
Education	0.83

These factors mean that software engineering is significantly different from the traditional medical arena in which systematic reviews were first developed.

Source: „Guidelines for performing Systematic Literature Reviews in SE“, Kitchenham et al., 2007

The review process

- Planning the review
- Conducting the review
- Reporting the review

Source: „Guidelines for performing Systematic Literature Reviews in SE“, Kitchenham et al., 2007

The review process

- Planning the review
  - Identification of the need for a review
  - (Commissioning a review)
  - Specifying the research questions
  - Developing a review protocol
  - (Evaluating the review protocol)
- Conducting the review
- Reporting the review

Source: „Guidelines for performing Systematic Literature Reviews in SE“, Kitchenham et al., 2007

The review process

- Planning the review
- Conducting the review
  - Identification of research
  - Selection of primary research
  - Study quality assessment
  - Data extraction and monitoring
  - Data synthesis
- Reporting the review

Source: „Guidelines for performing Systematic Literature Reviews in SE“, Kitchenham et al., 2007

The review process

- Planning the review
- Conducting the review
- Reporting the review
  - Specifying dissemination mechanisms
  - Formatting the main report
  - Evaluating the report

Source: „Guidelines for performing Systematic Literature Reviews in SE“, Kitchenham et al., 2007

The need for a review

Examples

Kitchenham et al. [21] argued that accurate cost estimation is important for the software industry; that accurate cost estimation models rely on past project data; that many companies cannot collect enough data to construct their own models. Thus, it is important to know whether models developed from data repositories can be used to predict costs in a specific company. They noted that a number of studies have addressed that issue but have come to different conclusions. They concluded that it is necessary to determine whether, or under what conditions, models derived from data repositories can support estimation in a specific company.

Jørgensen [17] pointed out in spite of the fact that most software cost estimation research concentrates on formal cost estimation models and that a large number of IT managers know about tools that implement formal models, most industrial cost estimation is based on expert judgement. He argued that researchers need to know whether software professionals are simply irrational, or whether expert judgement is just as accurate as formal models or has other advantages that make it more acceptable than formal models.

Source: „Guidelines for performing Systematic Literature Reviews in SE“, Kitchenham et al., 2007

The research questions

Specifying the research questions is the most important part of any systematic review. The review questions drive the entire systematic review methodology:

- The search process must identify primary studies that address the research questions.
- The data extraction process must extract the data items needed to answer the questions.
- The data analysis process must synthesise the data in such a way that the questions can be answered.

Source: „Guidelines for performing Systematic Literature Reviews in SE“, Kitchenham et al., 2007

Question types in health care

The Australian NHMR Guidelines [1] identify six types of health care questions that can be addressed by systematic reviews:

1. Assessing the effect of intervention.
2. Assessing the frequency or rate of a condition or disease.
3. Determining the performance of a diagnostic test.
4. Identifying aetiology and risk factors.
5. Identifying whether a condition can be predicted.
6. Assessing the economic value of an intervention or procedure.

Source: „Guidelines for performing Systematic Literature Reviews in SE“, Kitchenham et al., 2007

Question types in software engineering

In software engineering, it is not clear what the equivalent of a diagnostic test would be, but the other questions can be adapted to software engineering issues as follows:

- Assessing the effect of a software engineering technology.
- Assessing the frequency or rate of a project development factor such as the adoption of a technology, or the frequency or rate of project success or failure.
- Identifying cost and risk factors associated with a technology.
- Identifying the impact of technologies on reliability, performance and cost models.
- Cost benefit analysis of employing specific software development technologies or software applications.

Source: „Guidelines for performing Systematic Literature Reviews in SE“, Kitchenham et al., 2007

Asking the right question

The critical issue in any systematic review is to ask the right question. In this context, the right question is usually one that:

- Is meaningful and important to practitioners as well as researchers. For example, researchers might be interested in whether a specific analysis technique leads to a significantly more accurate estimate of remaining defects after design inspections. However, a practitioner might want to know whether adopting a specific analysis technique to predict remaining defects is more effective than expert opinion at identifying design documents that require re-inspection.
- Will lead either to changes in current software engineering practice or to increased confidence in the value of current practice. For example, researchers and practitioners would like to know under what conditions a project can safely adopt agile technologies and under what conditions it should not.
- Will identify discrepancies between commonly held beliefs and reality.

Source: „Guidelines for performing Systematic Literature Reviews in SE“, Kitchenham et al., 2007

Examples of research questions

Jørgensen [17] had two research questions:

1. Should we expect more accurate effort estimates when applying expert judgment or models?
2. When should software development effort estimates be based on expert judgment, when on models, and when on a combination of expert judgment and models?

Source: „Guidelines for performing Systematic Literature Reviews in SE“, Kitchenham et al., 2007

Question structure

Medical guidelines recommend considering a question about the effectiveness of a treatment from three viewpoints:

- The population, i.e. the people affected by the intervention.
- The interventions, which are usually a comparison between two or more alternative treatments.
- The outcomes, i.e. the clinical and economic factors that will be used to compare the interventions.

More recently Petticrew and Roberts suggest using the PICOC (Population, Intervention, Comparison, Outcome, Context) criteria to frame research questions [25]. These criteria extend the original medical guidelines with:
Comparison: I.e. what is the intervention being compared with
Context: i.e. what is the context in which the intervention is delivered.

Source: „Guidelines for performing Systematic Literature Reviews in SE“, Kitchenham et al., 2007

Question structure in *SE*

Population

In software engineering experiments, the populations might be any of the following:

- A specific software engineering role e.g. testers, managers.
- A category of software engineer, e.g. a novice or experienced engineer.
- An application area e.g. IT systems, command and control systems.
- An industry group such as Telecommunications companies, or Small IT companies.

A question may refer to very specific population groups e.g. novice testers, or experienced software architects working on IT systems. In medicine the populations are defined in order to reduce the number of prospective primary studies. In software engineering far fewer primary studies are undertaken, thus, we may need to avoid any restriction on the population until we come to consider the practical implications of the systematic review.

Source: „Guidelines for performing Systematic Literature Reviews in SE“, Kitchenham et al., 2007

Intervention

The intervention is the software methodology/tool/technology/procedure that addresses a specific issue, for example, technologies to perform specific tasks such as requirements specification, system testing, or software cost estimation.

Comparison

This is the software engineering methodology/tool/technology/procedure with which the intervention is being compared. When the comparison technology is the conventional or commonly-used technology, it is often referred to as the “control” treatment. The control situation must be adequately described. In particular “not using the intervention” is inadequate as a description of the control treatment. Software engineering techniques usually require training. If you compare people using a technique with people not using a technique, the effect of the technique is confounded with the effect of training. That is, any effect might be due to providing training not the specific technique. This is a particular problem if the participants are students.

Source: „Guidelines for performing Systematic Literature Reviews in SE“, Kitchenham et al., 2007

Outcomes

Outcomes should relate to factors of importance to practitioners such as improved reliability, reduced production costs, and reduced time to market. All relevant outcomes should be specified. For example, in some cases we require interventions that improve some aspect of software production without affecting another e.g. improved reliability with no increase in cost.

A particular problem for software engineering experiments is the widespread use of surrogate measures for example, defects found during system testing as a surrogate for quality, or coupling measures for design quality. Studies that use surrogate measures may be misleading and conclusions based on such studies may be less robust.

Source: „Guidelines for performing Systematic Literature Reviews in SE“, Kitchenham et al., 2007

Context

For Software Engineering, this is the context in which the comparison takes place (e.g. academia or industry), the participants taking part in the study (e.g. practitioners, academics, consultants, students), and the tasks being performed (e.g. small scale, large scale). Many software experiments take place in academia using student participants and small scale tasks. Such experiments are unlikely to be representative of what might occur with practitioners working in industry. Some systematic reviews might choose to exclude such experiments although in software engineering, these may be the only type of studies available.

Experimental designs

In medical studies, researchers may be able to restrict systematic reviews to primary studies of one particular type. For example, Cochrane reviews are usually restricted to randomised controlled trials (RCTs). In other circumstances, the nature of the question and the central issue being addressed may suggest that certain study designs are more appropriate than others. However, this approach can only be taken in a discipline where the large number of research papers is a major problem. In software engineering, the paucity of primary studies is more likely to be the problem for systematic reviews and we are more likely to need protocols for aggregating information from studies of widely different types.

Source: „Guidelines for performing Systematic Literature Reviews in SE“, Kitchenham et al., 2007

The review protocol

- Background. The rationale for the survey.
- The research questions that the review is intended to answer.
- The strategy that will be used to search for primary studies including search terms and resources to be searched. Resources include digital libraries, specific journals, and conference proceedings. An initial mapping study can help determine an appropriate strategy.
- Study selection criteria. Study selection criteria are used to determine which studies are included in, or excluded from, a systematic review. It is usually helpful to pilot the selection criteria on a subset of primary studies.
- Study selection procedures. The protocol should describe how the selection criteria will be applied e.g. how many assessors will evaluate each prospective primary study, and how disagreements among assessors will be resolved.
- Study quality assessment checklists and procedures. The researchers should develop quality checklists to assess the individual studies. The purpose of the quality assessment will guide the development of checklists.
- Data extraction strategy. This defines how the information required from each primary study will be obtained. If the data require manipulation or assumptions and inferences to be made, the protocol should specify an appropriate validation process.
- Synthesis of the extracted data. This defines the synthesis strategy. This should clarify whether or not a formal meta-analysis is intended and if so what techniques will be used.
- Dissemination strategy (if not already included in a commissioning document).
- Project timetable. This should define the review schedule.

Source: „Guidelines for performing Systematic Literature Reviews in SE“, Kitchenham et al., 2007

Identification of research

The aim of a systematic review is to find as many primary studies relating to the research question as possible using an unbiased search strategy. The rigour of the search process is one factor that distinguishes systematic reviews from traditional reviews.

It is necessary to determine and follow a search strategy. This should be developed in consultation with librarians or others with relevant experience. Search strategies are usually iterative and benefit from:

- Preliminary searches aimed at both identifying existing systematic reviews and assessing the volume of potentially relevant studies.
- Trial searches using various combinations of search terms derived from the research question.
- Checking trial research strings against lists of already known primary studies.
- Consultations with experts in the field.

Source: „Guidelines for performing Systematic Literature Reviews in SE“, Kitchenham et al., 2007

Publication bias

Publication bias refers to the problem that *positive* results are more likely to be published than *negative* results. The concept of *positive* or *negative* results sometimes depends on the viewpoint of the researcher. (For example, evidence that full mastectomies were not always required for breast cancer was actually an extremely positive result for breast cancer sufferers.)

Publication bias can lead to systematic bias in systematic reviews unless special efforts are made to address this problem. Many of the standard search strategies identified above are used to address this issue including:

- Scanning the grey literature
- Scanning conference proceedings
- Contacting experts and researchers working in the area and asking them if they know of any unpublished results.

In addition, statistical analysis techniques can be used to identify the potential significance of publication bias (see Section 6.5.7).

Source: „Guidelines for performing Systematic Literature Reviews in SE“, Kitchenham et al., 2007

Documenting the search

The process of performing a systematic literature review must be transparent and replicable (as far as possible):

- The review must be documented in sufficient detail for readers to be able to assess the thoroughness of the search.
- The search should be documented as it occurs and changes noted and justified.
- The unfiltered search results should be saved and retained for possible reanalysis.

Source: „Guidelines for performing Systematic Literature Reviews in SE“, Kitchenham et al., 2007

Documenting the search

Data Source	Documentation
Digital Library	Name of database Search strategy for the database Date of search Years covered by search
Journal Hand Searches	Name of journal Years searched Any issues not searched
Conference proceedings	Title of proceedings Name of conference (if different) Title translation (if necessary) Journal name (if published as part of a journal)
Efforts to identify unpublished studies	Research groups and researchers contacted (Names and contact details) Research web sites searched (Date and URL)
Other sources	Date Searched/Contacted URL Any specific conditions pertaining to the search

Source: „Guidelines for performing Systematic Literature Reviews in SE“, Kitchenham et al., 2007

Study selection criteria

Study selection criteria are intended to identify those primary studies that provide direct evidence about the research question. In order to reduce the likelihood of bias, selection criteria should be decided during the protocol definition, although they may be refined during the search process.
Inclusion and exclusion criteria should be based on the research question. They should be piloted to ensure that they can be reliably interpreted and that they classify studies correctly.

Example

Jørgensen [17] included papers that compare judgment-based and model-based software development effort estimation. He also excluded one relevant paper due to “incomplete information about how the estimates were derived”.

Source: „Guidelines for performing Systematic Literature Reviews in SE“, Kitchenham et al., 2007

More study selection criteria

- Language
- Journal
- Authors
- Setting
- Participants or subjects
- Research Design
- Sampling method
- Date of publication.

Source: „Guidelines for performing Systematic Literature Reviews in SE“, Kitchenham et al., 2007

Study quality assessment

In addition to general inclusion/exclusion criteria, it is considered critical to assess the "quality" of primary studies:

- To provide still more detailed inclusion/exclusion criteria.
- To investigate whether quality differences provide an explanation for differences in study results.
- As a means of weighting the importance of individual studies when results are being synthesised.
- To guide the interpretation of findings and determine the strength of inferences.
- To guide recommendations for further research.

An initial difficulty is that there is no agreed definition of study "quality". However, the CRD Guidelines [19] and the Cochrane Reviewers' Handbook [7] both suggest that quality relates to the extent to which the study minimises bias and maximises internal and external validity (see Table 3).

Source: „Guidelines for performing Systematic Literature Reviews in SE", Kitchenham et al., 2007

Quality concept definitions

Term	Synonyms	Definition
Bias	Systematic error	A tendency to produce results that depart systematically from the ‘true’ results. Unbiased results are internally valid
Internal validity	Validity	The extent to which the design and conduct of the study are likely to prevent systematic error. Internal validity is a prerequisite for external validity.
External validity	Generalisability, Applicability	The extent to which the effects observed in the study are applicable outside of the study.

Source: „Guidelines for performing Systematic Literature Reviews in SE“, Kitchenham et al., 2007

Example of quality assessment

Kitchenham et al. [21] constructed a quality questionnaire based on 5 issues affecting the quality of the study which were scored to provide an overall measure of *study* quality:

1. Is the data analysis process appropriate?
2. Did studies carry out a sensitivity or residual analysis?
3. Were accuracy statistics based on the raw data scale?
4. How good was the study comparison method?
5. The size of the within-company data set, measured according to the criteria presented below. Whenever a study used more than one within-company data set, the average score was used:

- Less than 10 projects: Poor quality (score = 0)
- Between 10 and 20 projects: Fair quality (score = 0.33)
- Between 21 and 40 projects: Good quality (score = 0.67)
- More than 40 projects: Excellent quality (score = 1)

Source: „Guidelines for performing Systematic Literature Reviews in SE“, Kitchenham et al., 2007

Data extraction forms

The data extraction forms must be designed to collect all the information needed to address the review questions and the study quality criteria. If the quality criteria are to be used to identify inclusion/exclusion criteria, they require separate forms (since the information must be collected prior to the main data extraction exercise). If the quality criteria are to be used as part of the data analysis, the quality criteria and the review data can be included in the same form.

In most cases, data extraction will define a set of numerical values that should be extracted for each study (e.g. number of subjects, treatment effect, confidence intervals, etc.). Numerical data are important for any attempt to summarise the results of a set of primary studies and are a prerequisite for meta-analysis (i.e. statistical techniques aimed at integrating the results of the primary studies).

Source: „Guidelines for performing Systematic Literature Reviews in SE“, Kitchenham et al., 2007

Example of a data extraction form

Data item	Value	Additional notes
Data Extractor		
Data Checker		
Study Identifier	S1	
Application domain	Space, military and industrial	
Name of database	European Space Agency (ESA)	
Number of projects in database (including within-company projects)	108	
Number of cross-company projects	60	
Number of projects in within-company data set	29	
Size metric(s): FP (Yes/No) Version used: LOC (Yes/No) Version used: Others (Yes/No) Number:	FP: No LOC: Yes (KLOC) Others: No	
Number of companies	37	
Number of countries represented	8	European only
Were quality controls applied to data collection?	No	
If quality control, please describe		
How was accuracy measured?	Measures: R^2 (for model construction only) MMRE Pred(25) r (Correlation between estimate and actual)	

Source: „Guidelines for performing Systematic Literature Reviews in SE“, Kitchenham et al., 2007

Cross-company model		
What technique(s) was used to construct the cross-company model?	A preliminary productivity analysis was used to identify factors for inclusion in the effort estimation model. Generalised linear models (using SAS). Multiplicative and Additive models were investigated. The multiplicative model is a logarithmic model.	
If several techniques were used which was most accurate?	In all cases, accuracy assessment was based on the logarithmic models not the additive models.	It can be assumed that linear models did not work well.
What transformations if any were used?	Not clear whether the variables were transformed or the GLM was used to construct a log-linear model	Not important: the log models were used and they were presented in the raw data form – thus any accuracy metrics were based on raw data predictions.
What variables were included in the cross-company model?	KLOC, Language subset, Category subset, RELY	Category is the type of application. RELY is reliability as defined by Boehm (1981)
What cross-validation method was used?	A hold-out sample of 9 projects from the single company was used to assess estimate accuracy	
Was the cross-company model compared to a baseline to check if it was better than chance?	Yes	The baseline was the correlation between the estimates and the actuals for the hold-out.
What was/were the measure(s) used as benchmark?	The correlation between the prediction and the actual for the single company was tested for statistical significance. (Note it was significantly different from zero for the 20 project data set, but not the 9 project hold-out data set.)	

Source: „Guidelines for performing Systematic Literature Reviews in SE“, Kitchenham et al., 2007

Within-company model		
What technique(s) was used to construct the within-company model?	A preliminary productivity analysis was used to identify factors for inclusion in the effort estimation model. Generalised linear models (using SAS). Multiplicative and Additive models were investigated. The multiplicative model is a logarithmic model.	
If several techniques were used which was most accurate?	In all cases, accuracy assessment was based on the logarithmic models not the additive models.	It can be assumed that linear models did not work well.
What transformations if any were used?	Not clear whether the variables were transformed or the GLM was used to construct a log-linear model	Not important: the log models were used and they were presented in the raw data form – thus any accuracy metrics were based on raw data predictions.
What variables were included in the within-company model?	KLOC, Language subset, Year	
What cross-validation	A hold-out sample of 9 projects	
method was used	from the single company was used to assess estimate accuracy	

Source: „Guidelines for performing Systematic Literature Reviews in SE“, Kitchenham et al., 2007

Data Summary		
Data base summary (all projects) for size and effort metrics.	Effort min: 7.8 MM Effort max: 4361 MM Effort mean: 284 MM Effort median: 93 MM Size min: 2000 KLOC Size max: 413000 KLOC Size mean: 51010 KLOC Size median: 22300 KLOC	KLOC: non-blank, non-comment delivered 1000 lines. For reused code Boehm's adjustment were made (Boehm, 1981). Effort was measured in man months, with 144 man hours per man month
With-company data summary for size and effort metrics.	Effort min: Effort max: Effort mean: Effort median: Size min: Size max: Size mean: Size median:	Not specified

Source: „Guidelines for performing Systematic Literature Reviews in SE“, Kitchenham et al., 2007

Data extraction procedures

Whenever feasible, data extraction should be performed independently by two or more researchers. Data from the researchers must be compared and disagreements resolved either by consensus among researchers or arbitration by an additional independent researcher. Uncertainties about any primary sources for which agreement cannot be reached should be investigated as part of any sensitivity analyses. A separate form must be used to mark and correct errors or disagreements.

If several researchers each review different primary studies because time or resource constraints prevent all primary papers being assessed by at least two researchers, it is important to employ some method of checking that researchers extract data in a consistent manner. For example, some papers should be reviewed by all researchers (e.g. a random sample of primary studies), so that inter-researcher consistency can be assessed.

For single researchers such as PhD students, other checking techniques must be used. For example supervisors could perform data extraction on a random sample of the primary studies and their results cross-checked with those of the student. Alternatively, a test-retest process can be used where the researcher performs a second extraction from a random selection of primary studies to check data extraction consistency.

Source: „Guidelines for performing Systematic Literature Reviews in SE“, Kitchenham et al., 2007

Descriptive (narrative) data synthesis

Extracted information about the studies (i.e. intervention, population, context, sample sizes, outcomes, study quality) should be tabulated in a manner consistent with the review question. Tables should be structured to highlight similarities and differences between study outcomes.

It is important to identify whether results from studies are consistent with one another (i.e. homogeneous) or inconsistent (e.g. heterogeneous). Results may be tabulated to display the impact of potential sources of heterogeneity, e.g. study type, study quality, and sample size.

Source: „Guidelines for performing Systematic Literature Reviews in SE“, Kitchenham et al., 2007

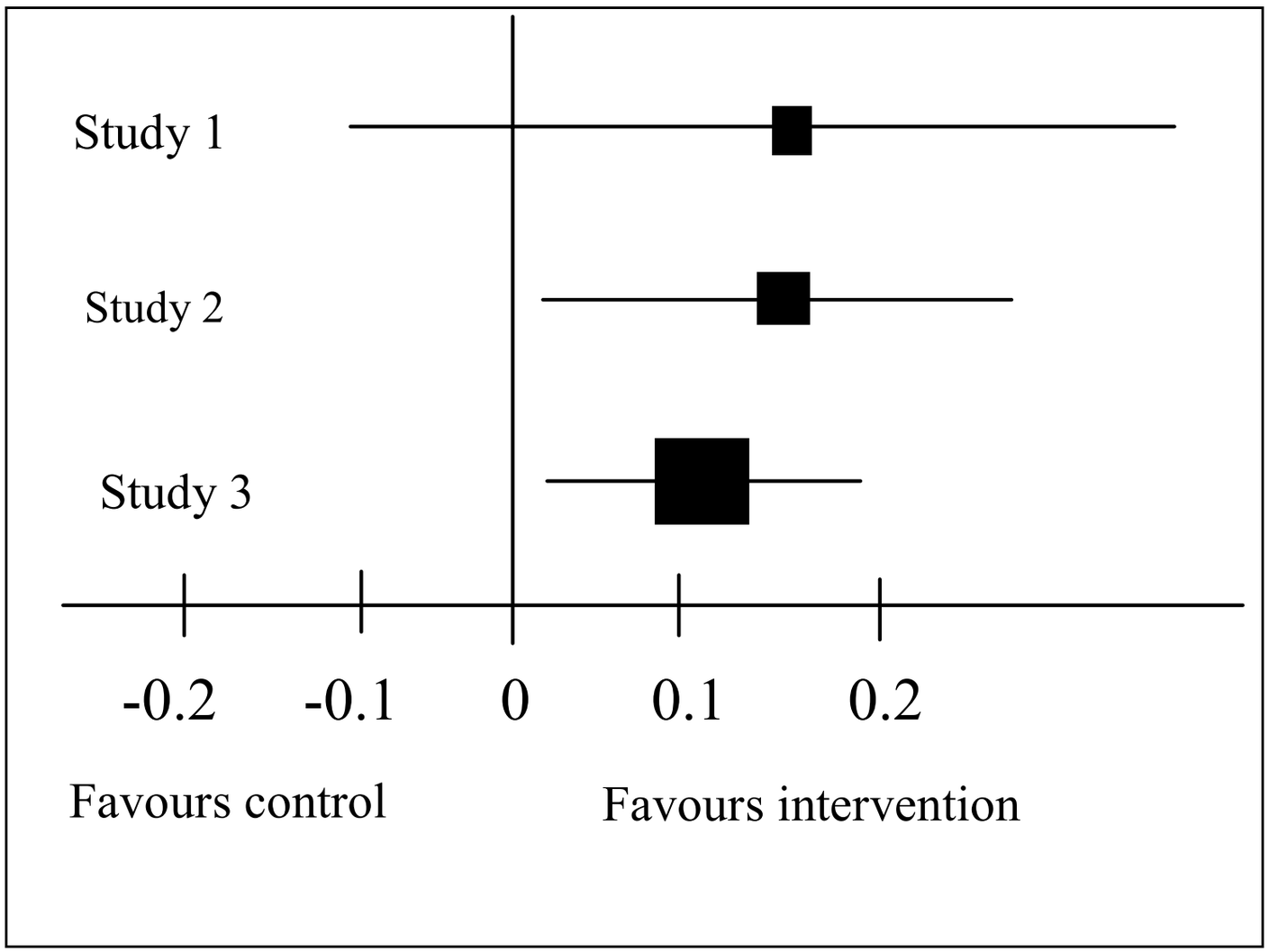
Quantitative data synthesis

Quantitative data should also be presented in tabular form including:

- Sample size for each intervention.
- Estimates effect size for each intervention with standard errors for each effect.
- Difference between the mean values for each intervention, and the confidence interval for the difference.
- Units used for measuring the effect.

Source: „Guidelines for performing Systematic Literature Reviews in SE“, Kitchenham et al., 2007

Example of presenting quantitative data as a forest plot



Source: „Guidelines for performing Systematic Literature Reviews in SE“, Kitchenham et al., 2007