SLR

Empirical Methods in Software Engineering (01OPJIU)

http://softeng.polito.it/EMSE/

Version 3.0

What is a SLR?

- Sport, Leicht, Rennsport

What is a SLR?

- Single-Lens Reflex

What is a SLR?

- Systematic
- Literature
- Review

COMPUTER NOVEMBER 2000
Information Retrieval//Component-Based Systems//Network Virtualization
COMPUTER JANUARY 2001
2001 Outlook//Nanoelectronics//Smart Dust//Automated Programming//IHMC//Interactive VR//Read
COMPUTER DECEMBER 2000
UML and SDL//Software Reliability//Vulnerability/Authentication//Cache-Concious Data Structures
Computer JULY 2005
Wi-Fi Network Security//Multiprocessor SoCs
Computer JUNE 2005
Technology in Developing Regions//Human Aspects of Software Engineering//Computing and Education
Misleading Architecting Tradeoffs//Virtualization Technologies
Computer MAY 2005

What

- A means of evaluating and interpreting all available studies relevant to a particular research question or phenomenon of interest.
- Systematic reviews aim to present a fair evaluation of a research topic by using a trustworthy, rigorous, and auditable methodology.

Evidence Based Medicine

Our vision is that healthcare decision-making throughout the world will be informed by high-quality, timely research evidence

http://www.cochrane.org

THE COCHRANE COLLABORATION®

Characteristics of SLR

- Clearly stated set of objectives with pre-defined eligibility criteria for studies
- Explicit, reproducible methodology
- Systematic search that attempts to identify all studies that would meet the eligibility criteria
- Assessment of the validity of the findings of the included studies,
    - for example through the assessment of risk of bias
- Systematic presentation, and synthesis, of the characteristics and findings of the included studies.

n-ary studies

- Individual studies analyzed in a systematic review are called primary studies
- A systematic review itself can be considered a secondary study.
- A systematic review that analyzes the information in SLR is a tertiary study.

SLR are fashionable in SE

- Jan 2004 – June 2008 → 54 SLRs
  - 1 per month
- Jul 2008 – Dec 2009 → 67 SLRs
  - 3.7 per month

Agenda

- Motivation
- Review process
  - Planning
  - Execution
  - Reporting
- Lessons learned

MOTIVATION

Why

- Most research starts, or should start, with a literature review of some sort.
- Unless a literature review is thorough and fair, it is of little scientific value.
- A SLR synthesizes existing work in a manner that is fair and seen to be fair.

How

- SLRs must be undertaken in accordance with a predefined search strategy, that must allow the completeness of the search to be assessed.
- Researchers performing a SLR 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.

What for

- To summarize the existing evidence concerning a treatment, technology
    - E.g. to summarize 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 order to appropriately position new research activities.
- To examine the extent to which empirical evidence supports/contradicts theoretical hypotheses, even to assist the generation of new hypotheses

Pros & Cons

- The well-defined methodology makes it more likely that the results of the literature are unbiased.
- They can provide information about the effects of some phenomenon across a wide range of settings and empirical methods.
- In case of quantitative studies, it is possible to combine data using meta-analytic techniques.

- The major disadvantage of systematic literature reviews is that they require considerably more effort than traditional literature reviews.

Key elements

- Review protocol
- Search strategy
- Documentation
- Explicit criteria
- Information specification

Key elements

- Review protocol
  - SLRs start by defining a review protocol that specifies the research question being addressed and the methods that will be used to perform the review.
- Search strategy
  - SLRs are based on a defined search strategy that aims to detect as much of the relevant literature as possible.
- Documentation
  - SLRs document their search strategy so that readers can access its rigorous and completeness.
- Explicit criteria
  - SLRs require explicit inclusion and exclusion criteria to assess each potential primary study.
- Information specification
  - SLRs specify the information to be obtained from each primary study including quality criteria to evaluate each primary study.

Other review types

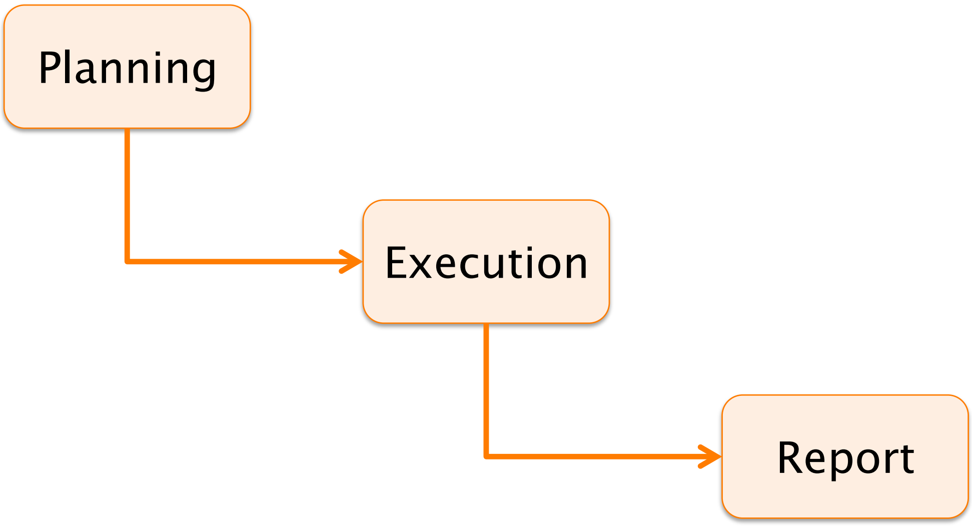
- Systematic Mapping Studies
  - When it is discovered that very little evidence is likely to exist, that the topic is very broad, then a systematic mapping study may be a more appropriate exercise than a systematic review.
  - A systematic mapping study allows the evidence in a domain to be plotted at a high level of granularity.
  - This allows for the identification of evidence clusters and evidence deserts to direct the focus of future systematic reviews and to identify areas for more primary studies to be conducted.

Other reviews

- Tertiary Reviews
  - In a domain where a number of systematic reviews exist already it may be possible to conduct a tertiary review, which is a systematic review of systematic reviews, in order to answer wider research questions.
  - A tertiary review uses exactly the same methodology as a standard systematic literature review.
  - It is potentially less resource intensive than conducting a new systematic review of primary studies but it is dependent on sufficient systematic reviews of a high quality being available.

REVIEW PROCESS

Review Process



Review process

- The stages may appear to be sequential, but it is important to recognize that many of the stages involve iteration.
- Many activities are initiated during the planning stage, and refined during execution. For example:
  - The inclusion and exclusion criteria are initially specified when the protocol is drafted but may be refined after quality criteria are defined.
  - Data extraction forms initially prepared during construction of the protocol will be amended when quality criteria are agreed.
  - Data synthesis methods defined in the protocol may be amended once data has been collected.

Running Examples

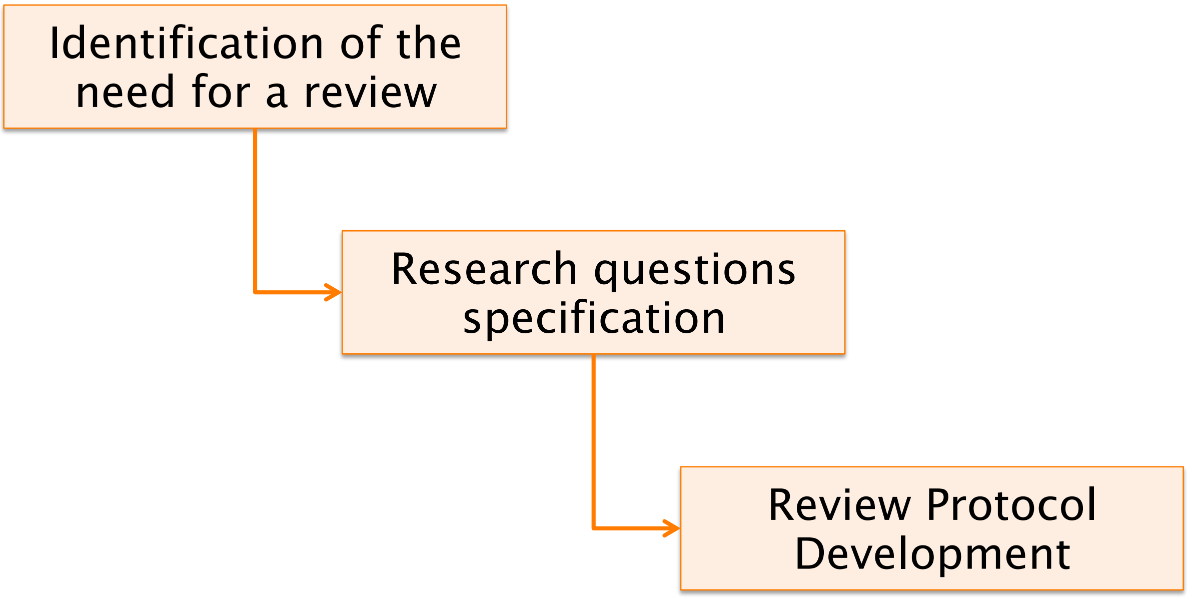
(Est1) Jorgensen, M., Shepperd, M. (2007). A systematic review of software development cost estimation studies. IEEE TSE 33(1), 33-53.

(Est2) Kitchenham, B., Mendes, E., Travassos, G.H. (2007) A Systematic Review of Cross- vs. Within-Company Cost Estimation Studies, IEEE TSE, 33 (5), 316-329.

(GSE) Darja Šmite , Claes Wohlin,Tony Gorschek, Robert Feldt. (2010). Empirical evidence in global software engineering: a systematic review. Empirical Software Engineering, 15, 91-118.

(UML) Marcela Genero, Ana M. Fernández, H. James Nelson, Geert Poels, Mario Piattini. (2011). A Systematic Literature Review on the Quality of UML Models. Journal of Database Management, 22(3), 46-70.

Planning


Identification of the need for a review

Research questions specification

Review Protocol Development

Identifying the need

- Researchers should identify and review any existing systematic reviews of the phenomenon of interest against appropriate evaluation criteria:
  - What are the review's objectives?
  - What sources were searched to identify primary studies? Were there any restrictions?
  - What were the inclusion/exclusion criteria and how were they applied?
  - What criteria were used to assess the quality of primary studies and how were they applied?
  - How were the data extracted from the primary studies?
  - How were the data synthesized? How were differences between studies investigated? How were the data combined? Was it reasonable to combine the studies? Do the conclusions flow from the evidence?

Commissioning

- When an organization requires information about a specific topic but does not have the time, expertise to perform a systematic literature itself.
- It will commission researchers to perform a systematic literature review of the topic.
- A commissioning document specifying the work required must be written.

Commissioning Document

- Project Title
- Background
- Review Questions
- Advisory/Steering Group (Researchers, Practitioners, Lay members, Policy Makers etc.)
- Methods for the review
- Project Timetable
- Dissemination Strategy
- Support Infrastructure
- Budget
- References

Research question

- Specifying the research questions is the most important part of any systematic review.
- The review questions drive the entire SLR 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 synthesize the data in such a way that the questions can be answered.

Research question

- The critical issue in any systematic review is to ask the right question.
- The right question is usually one that:
  - Is meaningful and important to practitioners as well as researchers.
  - Will lead either to changes in current software engineering practice, to increased confidence in the value of current practice.
  - Identify discrepancies between commonly held beliefs and reality.

EBSE recommendation

- How to appraise and apply methods, tools, and techniques in practice
- Three components:
  - The main intervention or action you're interested in
  - The context or specific situations of interest
  - The main outcomes or effects of interest

Cochrane

- To assess the effects of [*intervention or comparison*] for [*health problem*] in [*types of people, disease or problem and setting if specified*]
- Participants
- Intervention
- Comparison
- Outcomes

Question types

- Assessing the effect of a software engineering technology.
- Assessing the frequency/rate of a project development factor such as the adoption of a technology, the frequency/rate of project success, 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 software technologies.

RQ Example (GSE)

RQ1: What is the state-of-the-art in empirical studies of GSE?

RQ1.1: Who is Involved in GSE?

RQ1.2: Where are the Development Sites Located?

RQ1.3: What is Studied in GSE?

RQ1.4: How Successful are the Cases Reported in Literature?

RQ1.5: Why are Companies Involved in GSE?

RQ2: What is the strength of the empirical evidence reflected in the empirical GSE?

RQ Example (Est1)

RQ1: Which journals include papers on software cost estimation?

RQ2: How easy is it to identify relevant software cost estimation journal papers?

RQ3: To what extent are software cost estimation researchers aware of the breadth of potential estimation study sources?

RQ4: Which journal is the dominant software cost estimation journal? To what extent does this journal have research topic biases?

RQ5: How many researchers are there who have a long term interest in software cost estimation? To what extent do the interests of these researchers affect the distribution of research topics?

RQ Example (Est1)

RQ6: What are the most investigated software cost estimation research topics and how has this changed over time?

RQ7: What are the most investigated estimation methods and how has this changed over time?

RQ8: Is there a need for change of research focus?

RQ9: What are the most frequently applied research methods, and in what study context? How has this changed over time?

RQ10: What are the main shortcomings regarding use of research methods?

RQ Example (Est2)

RQ1: What evidence is there that cross-company estimation models are not significantly worse than within-company estimation models for predicting effort for software/Web projects?

RQ2: Do the characteristics of the study data sets and the data analysis methods used in the study affect the outcome of within-company and cross-company effort estimation accuracy studies?

RQ3: Which estimation method(s)/process(es) were best for constructing cross-company effort estimation models?

RQ Example (UML)

RQ1. Which type of UML model quality has been investigated by researchers?

RQ2. Which research methods are used in research on UML model quality?

RQ3. What is the nature of the research results on UML model quality?

RQ4. Which research goals are aimed at in research on UML model quality?

RQ5. Which type of UML diagrams is the focus of the research on UML model quality?

Review Protocol

- A review protocol specifies the methods that will be used to undertake specific systematic review, to reduce the possibility of researcher bias.
- The components of a protocol include all the elements of the review plus some additional planning information

Review Protocol

- Background, the rationale for the survey.
- The research questions that the review is intended to answer
- The search 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.
- 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.

Review Protocol

- **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.

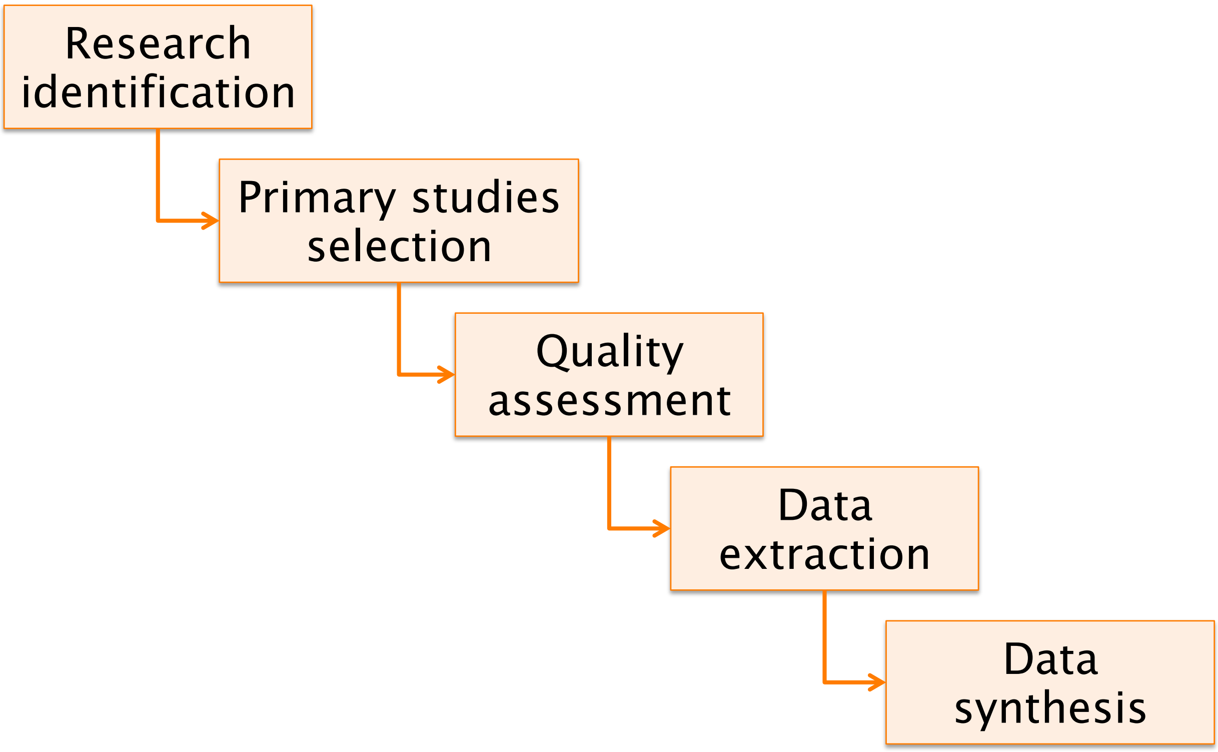
Review Protocol

- **Data extraction strategy**. This defines how the information required from each primary study will be obtained.
- **Synthesis of the extracted data**. This defines the synthesis strategy. This should clarify whether, 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.

Protocol Review

- The protocol is a critical element of any systematic review.
- Researchers must agree a procedure for reviewing the protocol.
- If appropriate funding is available, a group of independent experts should be asked to review the protocol.
- The same experts can later be asked to review the final report.
- PhD students should present their protocol to their supervisors for review and criticism.

Execution



Research identification

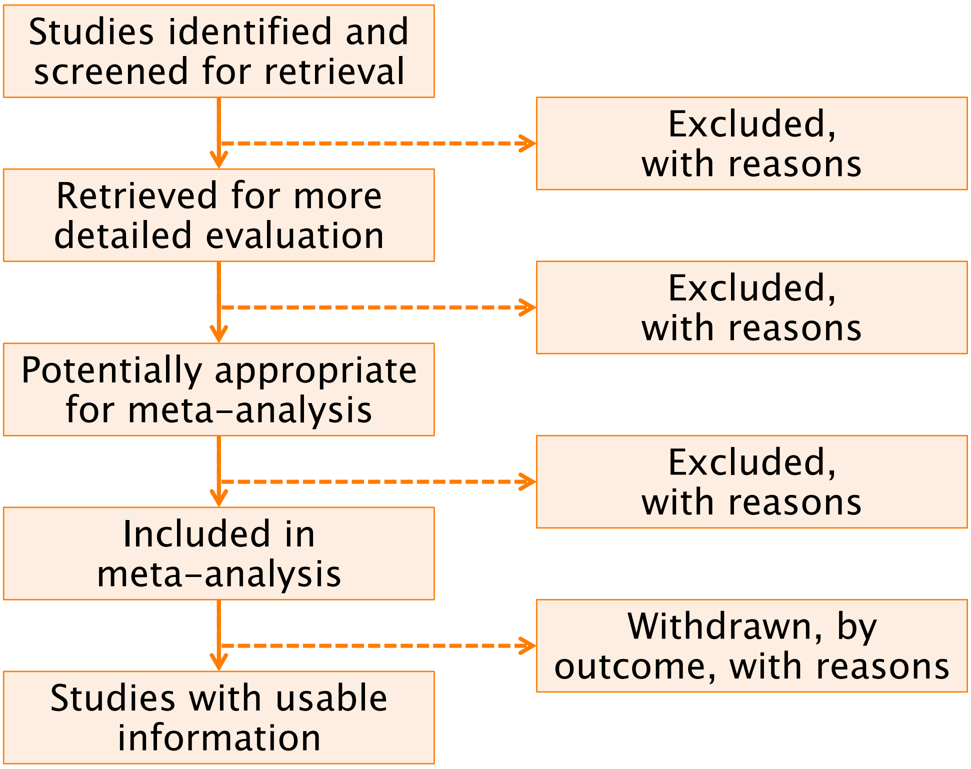
Primary studies selection

Quality assessment

Data extraction

Data synthesis

Stages and measures



Studies identified and screened for retrieval

Excluded, with reasons

Retrieved for more detailed evaluation

Excluded, with reasons

Potentially appropriate for meta-analysis

Excluded, with reasons

Included in meta-analysis

Withdrawn, by outcome, with reasons

Studies with usable information

Research identification

- Search strategy
  - Search strings
- Publication bias
- Bibliography management and document retrieval
- Documenting the search

Search strategy

- 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 searchers using various combinations of search terms derived from the research question
  - Reviews of research results
  - Consultations with experts in the field

Search strategy

- Initial searches for primary studies can be undertaken initially using electronic databases but this is not sufficient.
- Other sources of evidence must also be searched (sometimes manually) including:
  - Reference lists from relevant primary studies and review articles
  - Journals (including company journals such as the IBM Journal of Research and Development), grey literature (i.e. technical reports, work in progress) and conference proceedings
  - Research registers
  - The Internet

Search string

- Constructed using the following steps:
  - Define the major terms
  - Identify alternative spellings, synonyms, related terms for major terms.
  - Check the keywords in any relevant papers we already had.
  - Use the Boolean, to incorporate alternative spellings, synonyms, related terms.
  - Use the Boolean AND to link the major terms

Search string – Example (UML)

- Major: Quality
  - Synonyms and related: consistency, maintainability, understandability, completeness, comprehension, comprehensibility, testability, defect, effectiveness, complexity, readability, metric, measure, efficiency, validation, verification, layout
- Major: UML
  - Synonyms and related: Unified Modeling Language
- Major: Representation
  - Synonyms and related: Representation, diagram, model

Search string – Example (UML)

- Resulting search string

(UML *OR* UNIFIED MODELING LANGUAGE)

AND

(REPRESENTATION *OR* DIAGRAM OR MODEL)

AND

(QUALITY *OR* CONSISTENCY *OR* MAINTAINABILITY *OR* UNDERSTANDABILITY *OR* COMPLETENESS *OR* COMPREHENSION *OR* COMPREHENSABILITY *OR* TESTABILITY *OR* DEFECT *OR* EFFECTIVENNES *OR* COMPLEXITY *OR* READABILITY *OR* EFFICIENCY *OR* VALIDATION *OR* VERIFICATION *OR* LAYOUT)

Search string – Example (Est2)

(software OR application OR product OR Web OR WWW OR Internet OR World-Wide Web OR project OR development)

AND

(method OR process OR system OR technique OR methodology OR procedure)

AND

(cross company OR cross organisation OR cross organization OR cross organizational OR cross organisational OR cross- company OR cross-organisation OR cross-organization OR cross-organizational OR cross-organisational OR multi company OR multi organisation OR multi organization OR multi organizational OR multi organisational OR multi- company OR multi-organisation OR multi-organization OR multi-organizational OR multi-organisational OR multiple company OR multiple organisation OR multiple organization OR multiple organizational OR multiple organisational OR multiple-company OR multiple-organisation OR multiple-organization OR multiple-organizational OR multiple- organisational OR within company OR within organisation OR within organization OR within organizational OR within organisational OR within-company OR within-organisation OR within-organization OR within-organizational OR within-organisational OR single company OR single organisation OR single organization OR single organizational OR single organisational OR single-company OR single-organisation OR single-organization OR single-organizational OR single-organisational OR company-specific)

AND

(model OR modeling OR modelling)

AND

(effort OR cost OR resource)

AND

(estimation OR prediction OR assessment)

Search string – Example (Est2)

```
(software OR application OR product OR Web OR WWW OR
Internet OR World-Wide Web OR project OR development)
AND
(method OR process OR system OR technique OR methodology
OR procedure)
AND
(cross company OR cross organisation OR cross
organization OR cross organizational OR … OR multiple
organizational OR multiple organisational OR …)
AND
(model OR modeling OR modelling)
AND
(effort OR cost OR resource)
AND
(estimation OR prediction OR assessment)
```

Search string – Example (GSE)

- The final search strings were based on the experience from the pilot searches and consisted of a Boolean expression:

  (A1 *or* A2 *or* A3 *or* A4)

  AND (B1 *or* B2 *or* B3 *or* B4)

- where
  - A1= global software development
  - A2= global software engineering
  - A3= distributed software development
  - A4= distributed software engineering
  - B1= empirical
  - B2= industrial
  - B3= experiment
  - B4= case study

Search source – Examples (UML)

- SCOPUS database,
- Science@Direct with the subject Computer Science,
- Wiley InterScience with the subject of Computer Science, I
- IEEE Digital Library,
- ACM Digital Library,
- SPRINGER database.

Search source – Examples (Est2)

- Sources
  - INSPEC
  - El Compendex
  - Science Direct
  - Web of Science
  - IEEExplore
  - ACM Digital library
- The search strings needed to be adapted to suit the specific requirements of the difference data bases.
- In addition, the researchers searched several individual journals (J) and conference proceedings (C) sources

Search sources – Example (GSE)

- Compendex,
- IEEE Xplore,
- Springer Link,
- ISI Web of Knowledge,
- ScienceDirect,
- Wiley Inter Science Journal Finder,
- ACM Digital Library

Publication bias

A.k.a. “the file drawer problem”

- The editorial predilection for publishing particular findings e.g., positive results, which leads to the failure of authors to submit negative findings for publication

R. Rosenthal (1979) The "file drawer problem" and tolerance for null results, *Psychological Bulletin,* Vol. 86, No. 3, 838-641

Publication bias

- Roughly 90% of the published studies confirm the experimental hypotheses being tested

⇩

- Only well designed studies, with high power are performed
- Researchers formulate only true hypotheses

Publication bias

- Different causes
  - Study design or execution
  - Researcher decision
  - Journals behavior
  - Sponsorship
  - Review design or execution

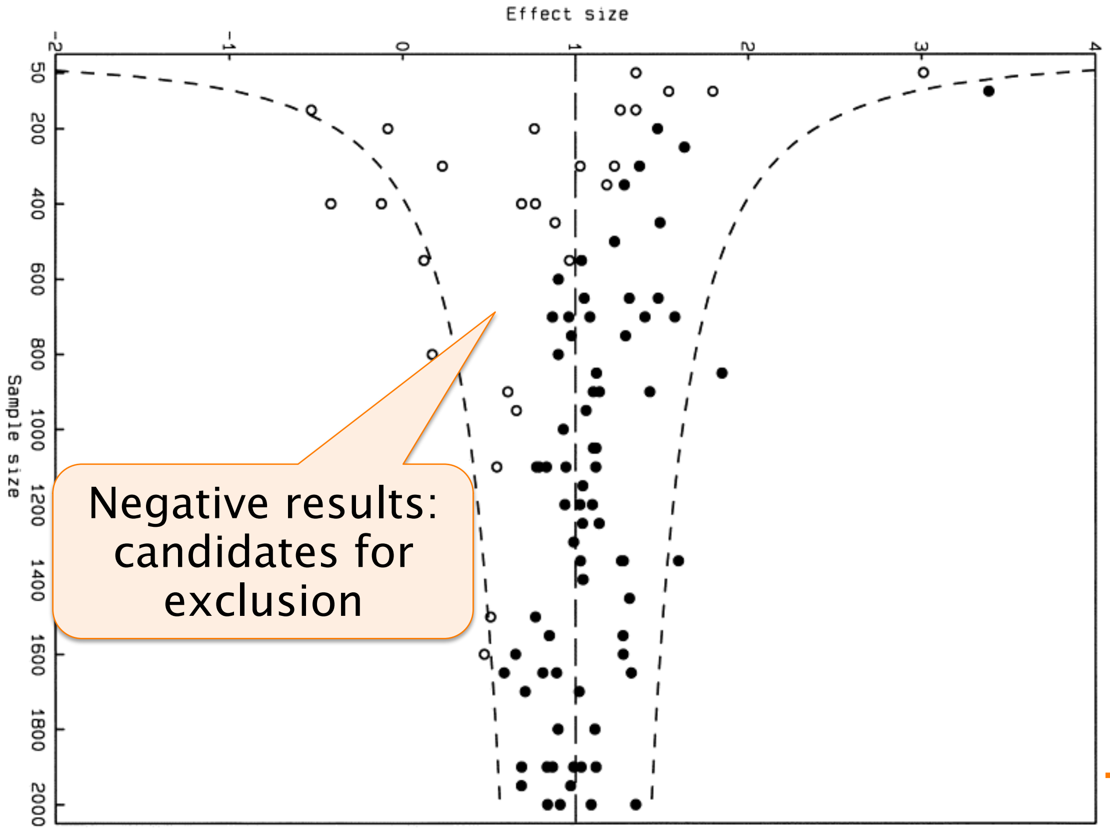
Publication bias

- Scanning the grey literature,
- Scanning conference proceedings.
- Contacting experts and researches working in the area and asking them if they know of any unpublished results
- Research registries
  - In medicine: ClinicalTrials.gov

Publication bias detection

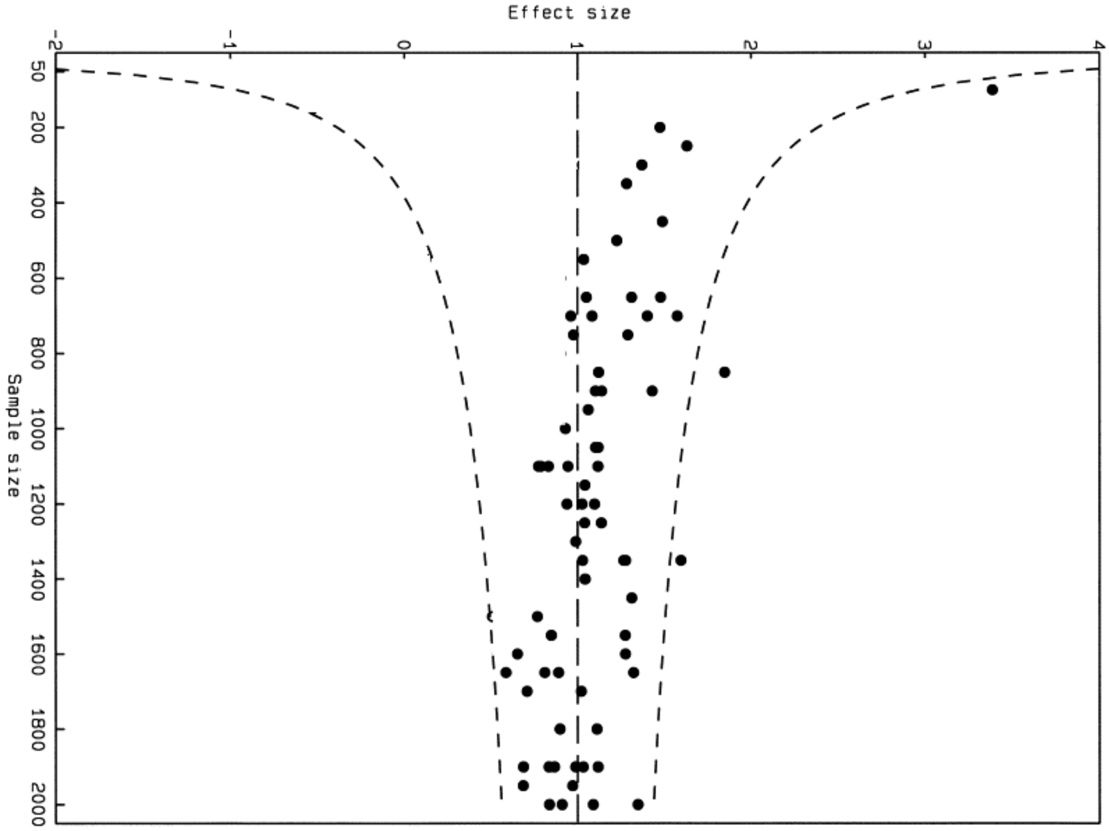
- Proportion of significant studies
  - Very simple
  - Does not actually demonstrate publication bias as no expected percentage of positive studies exists
- Funnel graphs
  - Only requires published data
  - Symmetry defined informally
- Statistical methods

Funnel plot



Negative results: candidates for exclusion

Funnel plot



Bib management and retrieval

- Bibliographic packages to manage the large number of references that can be obtained from a thorough literature research
  - Reference Manager,
  - Endnote
  - BibDesk
- Once reference lists have been finalized the full articles of potentially useful studies will need to be obtained.

Documenting the search

- The process of performing a systematic review must be transparent and replicable:
  - 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.

Documenting the search

Data source	Documentation items
Electronic library	Name of the database Search strategy Date of search Years covered by search
Journal (manual search)	Name of journal Years searched Any issue non searched
Conference proceedings	Title of proceedings (Name of conference) (Title translation)
Efforts to identify unpublished research	Research groups and researchers contacted Research web sites searched
Other sources	Date searched / contacted URL Any specific condition pertaining the search

Study selection

- Selection criteria should be decided during the protocol definition.
- 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.

Study selection

- Study selection process
  - Initially, selection criteria should be interpreted liberally, so that unless studies can be clearly excluded based on titles and abstracts, full copies should be obtained.
  - Final inclusion/exclusion decisions should be made after the full texts have been retrieved.
  - Maintain a list of excluded studies identifying the reason for exclusion.
- Reliability of inclusion decisions
  - When two or more researchers assess each paper, agreement between researchers must be reached

Selection criteria – Example(UML)

- Inclusion criteria:
  - Papers which dealt with UML and the tangible results of the modeling process (the UML diagram),
  - were written in English,
  - and were published between 1997 and 2009.

Selection criteria – Example(UML)

- Exclusion criteria:
  - pure discussion and opinion papers,
  - studies available only in the form of abstracts or PowerPoint presentations,
  - duplicates (for example, the same paper included in more than one database or in more than one journal),
  - research focusing issues other than UML model quality (for example, functional size measurement),
  - or where quality is mentioned only as a general introductory term in the paper’s abstract and an approach or other type of proposal related to quality is not amongst the paper’s contributions.
  - Papers were also excluded if they dealt with the quality and complexity of UML as a language (for example, how to make UML the language simpler) rather than on the quality and complexity of the models produced by UML,
  - if the paper was a summary of a workshop.

Selection Criteria – Example(Est2)

- Inclusion criteria:
  - any study that compared predictions of cross-company models with within-company models based on analysis of single company project data.
- Exclusion criteria:
  - studies where projects were only collected from a small number of different sources (e.g. 2 or 3 companies),
  - studies where models derived from a within-company data set were compared with predictions from a general cost estimation model.

Selection criteria – Example (Est1)

- Inclusion criteria
  - papers that compare judgment-based and model-based software development effort estimation.
- Exclusion criteria
  - excluded one relevant paper due to “incomplete information about how the estimates were derived”.

Snowball Sampling

- Establish a start set of papers
  - search in a general purpose engine (e.g. Google Scholar)
- Apply inclusion and exclusion criteria
- Then snowballing is performed
  - backward snowballing (based on reference lists)
  - forward snowballing (based on citations)

Study quality assessment

- It is generally considered important 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 synthesized.
  - To guide the interpretation of findings and determine the strength of inferences.
  - To guide recommendations for further research.

Quality assessment

- Quality relates to the extent to which the study minimizes bias and maximizes internal and external validity

Parameter	Synonyms	Description
Bias	Systematic error	Tendency to produce results that depart systematically from “true” results.
Internal validity	Validity	The extent to which design and conduct of the study are likely to prevent systematic error.
External validity	Generalizability, Applicability	The extent to which the effects observed in the study are applicable outside the study.

Quality assessment

- It is advisable to :
  - build checklists
  - assign numerical scales
    - quantitative assessments of quality can be obtained.
- Checklists are also developed by considering bias and validity problems that can occur at the different stages in an empirical study:
  - Design, Conduct, Analysis, and Conclusions.
- Kitchenham et al (2007) in the technical report provide:
  - A quality checklist for quantitative studies
  - A quality checklist for qualitative studies

Quality – Example(Est2)

- Is the data analysis process appropriate?
  - Was the data investigated to identify outliers and to assess distributional properties before analysis?
  - Was the result of the investigation used appropriately to transform the data and select appropriate data points?
- Did studies carry out a sensitivity or residual analysis?
  - Were the resulting estimation models subject to sensitivity or residual analysis?
  - Was the result of the sensitivity or residual analysis used to remove abnormal data points if necessary?
- Were accuracy statistics based on the raw data scale?
- How good was the study comparison method?
  - Was the single company selected at random (not selected for convenience) from several different companies?
  - Was the comparison based on an independent hold out sample (0.5) or random subsets (0.33), leave-one-out (0.17), no hold out (0)?

Quality – Example (Est2)

- 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)

Quality assessment

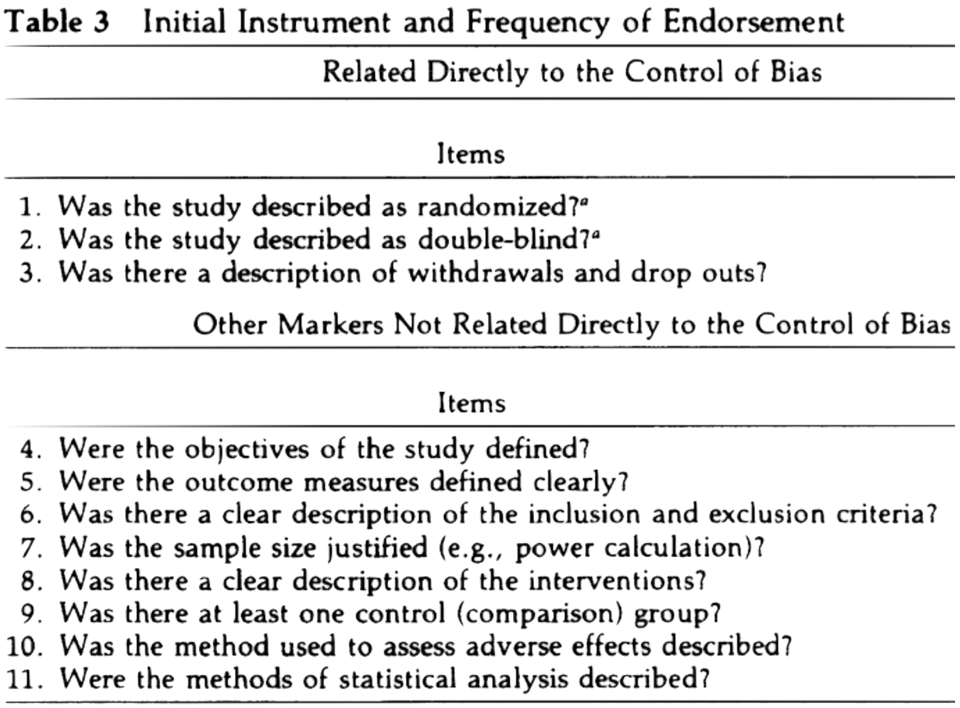


Table 3 Initial Instrument and Frequency of Endorsement

Related Directly to the Control of Bias

Items

1. Was the study described as randomized?[a]
2. Was the study described as double-blind?[a]
3. Was there a description of withdrawals and drop outs?

Other Markers Not Related Directly to the Control of Bias

Items

4. Were the objectives of the study defined?
5. Were the outcome measures defined clearly?
6. Was there a clear description of the inclusion and exclusion criteria?
7. Was the sample size justified (e.g., power calculation)?
8. Was there a clear description of the interventions?
9. Was there at least one control (comparison) group?
10. Was the method used to assess adverse effects described?
11. Were the methods of statistical analysis described?

Jadad et al., Assessing the quality of reports of randomized clinical trials: Is blinding necessary?, Controlled Clinical Trials, 17(1), February 1996

Data extraction

- Data extraction forms
  - All the questions needed to answer the review question
  - Quality evaluation criteria,
  - Standard information including:
    - Name of Review
    - Date of Data extraction
    - Title, authors, journal, publication details
    - Space for additional notes
- Data extraction procedures
- Multiple publications of the same data

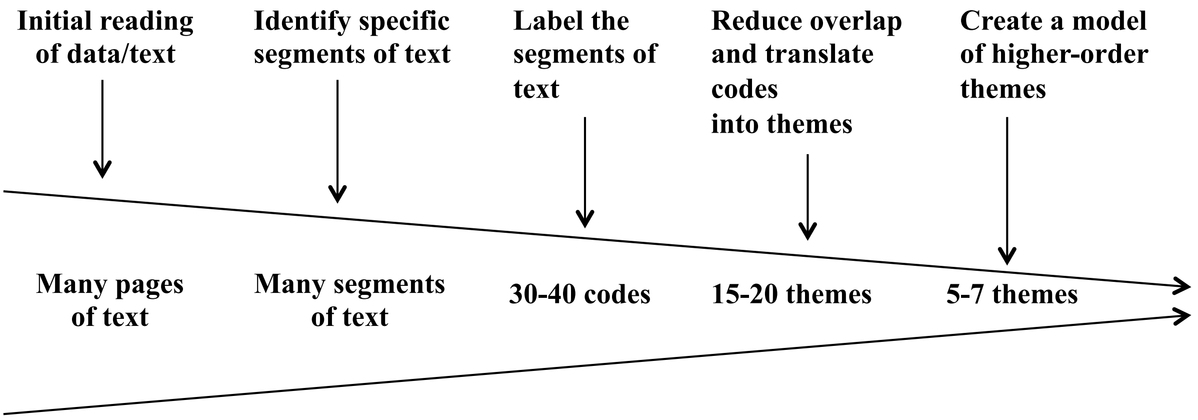
Data synthesis

- Descriptive synthesis (narrative)
  - Extracted information should be tabulated
- Qualitative synthesis
  - Thematic synthesis
  - Grounded theory
- Quantitative synthesis
  - Descriptive statistics
  - Meta-analysis

Thematic Synthesis

- Qualitative research method
- Well-organized way of describing large and potentially diverse evidence
- Generate new insights from primary studies, e.g. relating to rare or infrequent events

Thematic Synthesis



Initial reading of data/text

Identify specific segments of text

Label the segments of text

Reduce overlap and translate codes into themes

Create a model of higher-order themes

Many pages of text

Many segments of text

30-40 codes

15-20 themes

5-7 themes

From: D. S. Cruzes and T. Dyba, "Recommended Steps for Thematic Synthesis in Software Engineering," *2011 International Symposium on Empirical Software Engineering and Measurement*, 2011, pp.275-284.

TS – Extract Data

- Extract data from the primary studies, including bibliographical information, aims, context, and results.
  - Have all papers been read carefully to get immersed with the data?
  - Have specific segments of text pertaining to the objectives of the synthesis been identified?
  - Have publication details, context descriptions, and findings been extracted from all papers?
  - Have another researcher checked the extraction?

TS – Code Data

- Identify and code interesting concepts, categories, findings, and results in a systematic fashion across the entire data set.
  - Have important segments of text like concepts, categories, findings, and results been labeled and coded?
  - Has coding been done across the entire data set on a level that is appropriate for the research questions?
  - Has a list of initial codes with definitions and frequencies been created and checked by another researcher?
  - Have consistency checks or inter-rater reliability checks been performed to establish the credibility of the coding?
  - Are there clear, evident connections between the text and the codes?

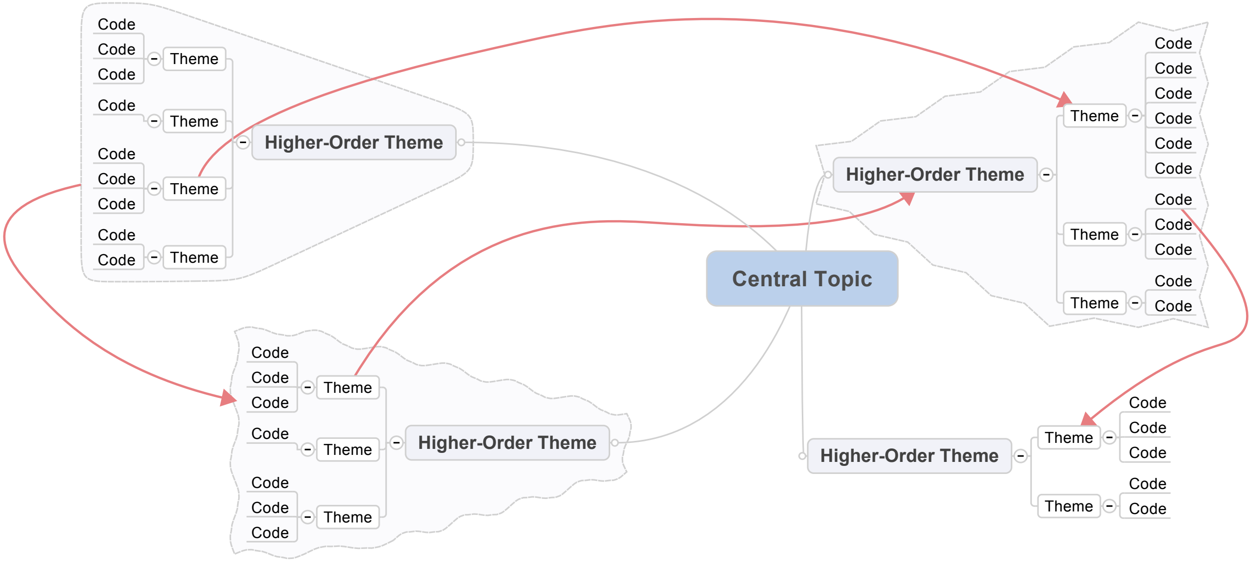
TS - Codes into themes

- Translate codes into themes, sub-themes, and higher order themes.
  - Have themes been created from a thorough, inclusive, and comprehensive review of the codes of all papers?
  - Has overlap between codes been reduced and the remaining codes been collated and translated into themes
  - Have themes been checked against each other and back to the data of the original papers?
  - Are themes internally coherent, consistent, and distinctive?

TS-Model of higher-order themes

- Explore relationships between themes and create a model of higher-order themes.
  - Have themes been compared across studies, translated into each other, and interpreted into higher-order themes?
  - Have higher-order themes and relationships between themes been checked against the research questions of the synthesis?
  - Are there clear descriptions of the higher-order themes and the relationships between these themes?
  - Has a model been created to show the relationships between the higher-order themes?

Themes



From: D. S. Cruzes and T. Dyba, "Recommended Steps for Thematic Synthesis in Software Engineering," *2011 International Symposium on Empirical Software Engineering and Measurement*, 2011, pp.275-284.

TS-Assess the trustworthiness

- Assess the trustworthiness of the interpretations leading up to the thematic synthesis.
  - Have the assumptions about, and specific approach to, the thematic synthesis been clearly explicated?
  - Is there a good fit between what is claimed and what the evidence shows?
  - Are the language and concepts used in the synthesis consistent?
  - Are there search questions answered based on the evidence of the thematic synthesis?

Qualitative analysis tool

- Tools for qualitative data analysis support the coding phase
  - Nvivo
  - Atlas.ti
  - QDA Miner

Meta analysis

- Statistical technique for combining the findings from independent studies.
- Often used to assess the practical effectiveness of methods and techniques;
  - two or more randomized controlled trials.
- Provides a precise estimate of treatment effect, giving due weight to the size of the different studies included.
- Good meta-analyses
  - aim for complete coverage of all relevant studies,
  - look for the presence of heterogeneity, and
  - explore the robustness of the main findings using sensitivity analysis.

Meta analysis

- Two stage process
  - summary statistic is calculated for each study, to describe the observed intervention effect
  - a summary (pooled) intervention effect estimate is calculated as a weighted average of the intervention effects estimated in the individual studies

$$summary = \frac{\text{sum of (estimante weight)}}{\text{sum of weights}} = \frac{\sum Y_i Wi}{\sum W_i}$$

Effect size

- The findings from individual studies are combined using an appropriate statistical method.
  - Separate methods are used for combining different outcome measures.
  - The methods use a similar approach in which the estimate from each study is weighted by the precision of the estimate.

Outcome measures

- Separate methods are used for combining different outcome measures
- Effect size
  - Dichotomous
  - Continuous
  - Ordinal
  - Count and rates
  - Time-to-event

Effect size – Continuous

- Mean difference
  - Difference of means
- Standardized mean difference
  - Cohen’s d
  - Hedges' *g*

$$SMD = \frac{\text{Difference of outcome means}}{\text{Standard deviation of outcome}} = \frac{\bar{x}_E - \bar{x}_C}{\boxed{s}}$$

Effect size – Dichotomous

- Success ratio
  - Odds-ratio
  - Risk ratio (Relative risk)
  - Same interpretation
    - E.g. ratio of 2 ➔defined outcome happens about twice as often in the intervention group as in the control group;
    - ratio of 0.5 ➔ around a 50% reduction in the defined event in the treated group compared with the controls

Ratios

	Event (Success)	No Event (Failure)	Marginals
Experimental intervention	S_E	F_E	N_E
Control Intervention	S_C	F_C	N_C
Marginals	N_S	N_F	N

$$RR = \frac{\text{risk of event in experimental group}}{\text{risk of event in control group}} = \frac{S_E/N_E}{S_C/N_C}$$

$$OR = \frac{\text{odds of event in experimental group}}{\text{odds of event in control group}} = \frac{S_E/F_E}{S_C/F_C} = \frac{S_E F_C}{F_E S_C}$$

Effect size - Rate

- Rates relate the counts of events to the amount of time during which they could have happened.
  - Counts of rare events (Poisson data)
- Rate-ratio (RR): which compares the rate of events in the two groups by dividing one by the other.

Meta-analysis models

- Fixed-effect
  - each study is estimating exactly the same quantity
- Random effect
  - studies are not all estimating the same intervention effect

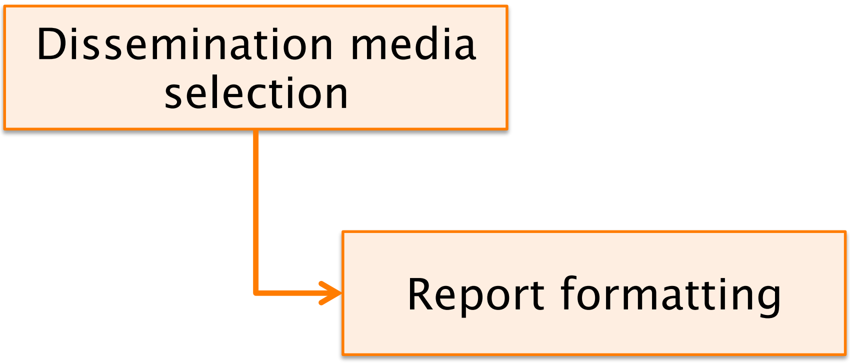
Methods

- Continuous
  - Inverse variance: $W_i = \frac{1}{SE_i^2}$
- Dichotomous
  - Mantel–Haenszel
    - Odds ratio $W_i = \frac{F_{E,i} \cdot S_{C,i}}{N_i}$
    - Risk ratio $W_i = \frac{S_{C,i}(S_{E,i} + F_{E,i})}{N_i}$
- Vote counting
  - number of positive vs. number of negative studies

Sensitivity analysis

- Explores ways in which the main findings are changed by varying:
  - Selection
  - Inclusion
  - Aggregation
  - Etc.
- Sort of threats to validity

Report


Dissemination media selection

Report formatting

Dissemination strategy

- It is important to communicate the results of a systematic review effectively.
- Most guidelines recommend planning the dissemination strategy during the commissioning stage (if any) or when preparing the systematic review protocol.
- Academics usually assume that dissemination is about reporting results in academic journals and/or conferences.

Dissemination Venues

- Journals
  - Information and Software Technology
  - IEEE Transactions on Software Engineering
  - Empirical Software Engineering
  - IEEE Software – Voice of Evidence column
- Conferences
  - ESEM
  - EASE

Dissemination strategy

- If the results of a systematic review are intended to influence practitioners, other forms of dissemination are necessary:
  - Practitioner journals and magazines,
  - Press releases to popular and specialized press,
  - Short summary leaflets,
  - Posters,
  - Web pages,
  - Direct communication to affected bodies.

Report formatting

- Typical formats:
  - Technical report or section of a PhD thesis.
  - Journal or conference paper.
- A journal or conference paper will normally have a size restriction.
  - In order to ensure that readers are able to properly evaluate the rigor and validity of a systematic review, journal papers should reference a technical report or thesis that contains all the details.

Report Structure

- Structured abstract
- Background
- Research questions
- Method
- Included and excluded studies
- Results
- Discussion
- Conclusions

Structured abstract

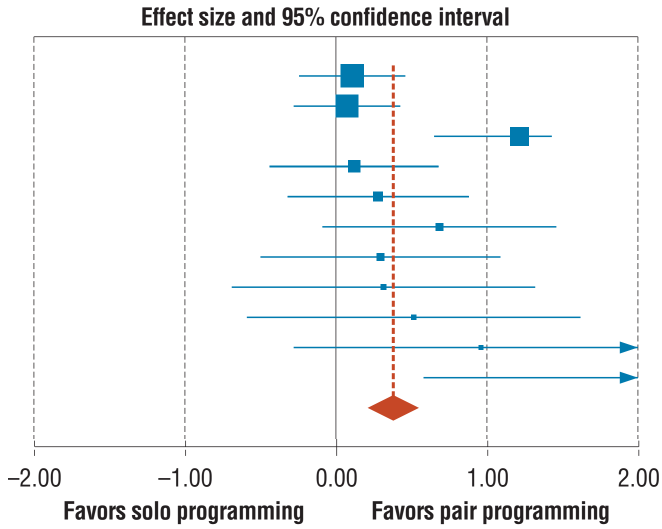
- Context
  - The important of the research questions addressed
- Objectives
  - The question addressed
- Methods
  - Data sources, study selection, quality assessment and data extraction
- Results
  - Main findings including meta-analysis results , and sensitivity analysis
- Conclusions
  - Implications for practice and future research

Graphical representation

- Forest plot
- Bubble plot (?)
- L’Abbé plot
- Galbraith (radial) plots

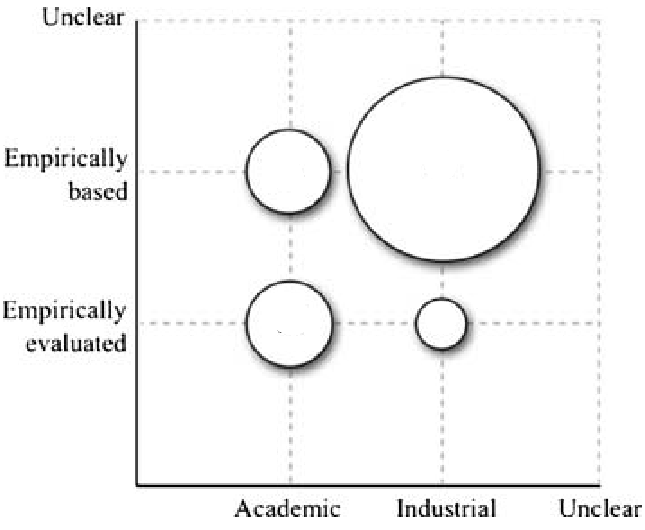
Forest plot

Study	Effect size	Lower limit	Upper limit	Relative weight
P07a	0.11	−0.24	0.46	23.24
S06a	0.08	−0.28	0.43	22.85
S00	1.04	0.65	1.43	18.64
S03	0.10	−0.44	0.64	9.80
S05b	0.28	−0.32	0.88	7.95
P07b	0.69	−0.09	1.46	4.73
S02	0.30	−0.50	1.09	4.53
S06c	0.32	−0.69	1.32	2.81
S06b	0.51	−0.59	1.62	2.33
P98	0.91	−0.28	2.10	2.02
S06d	2.20	0.58	3.82	1.09
Overall effect	**0.38**	**0.21**	**0.55**	



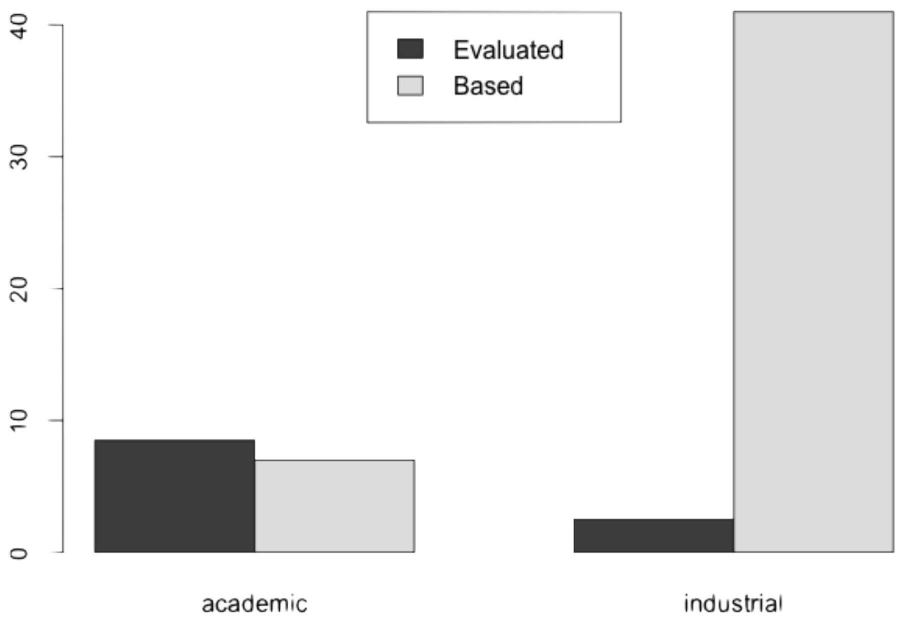
Dybå et al., 2007

Bubble plot

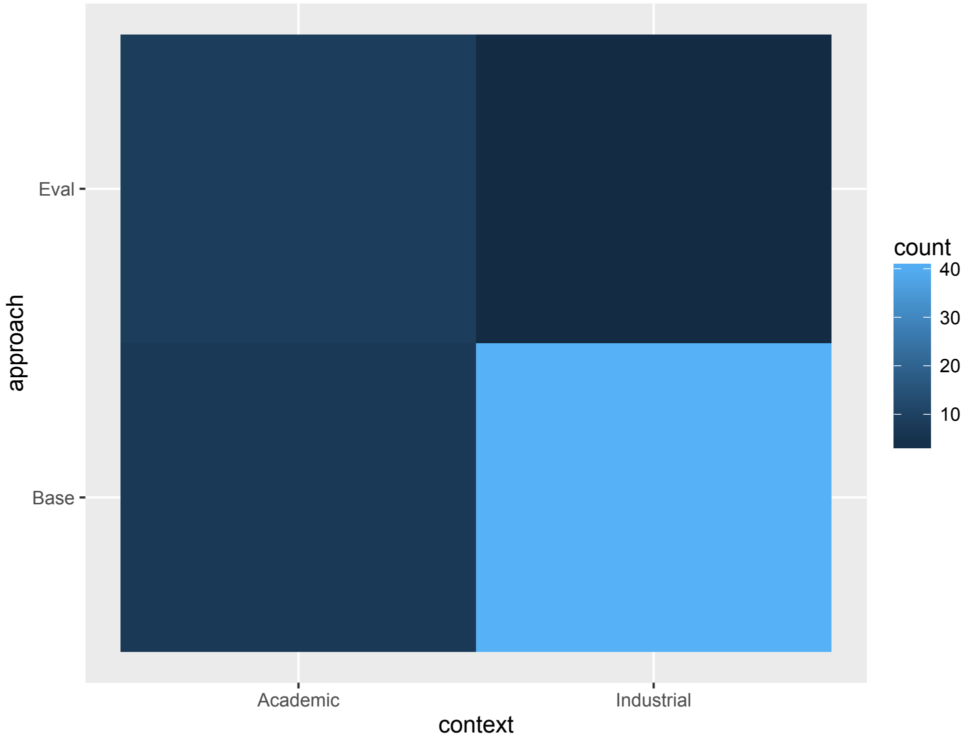


(GSE)

.. Or bar plot



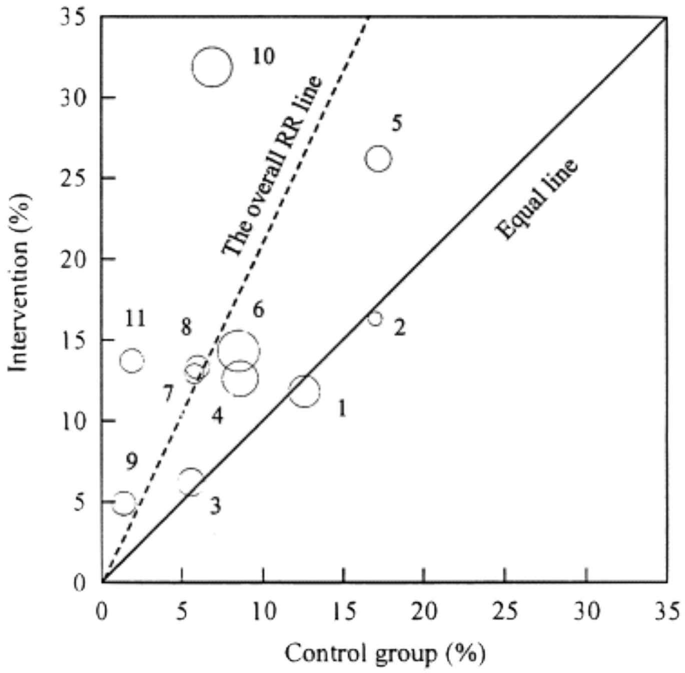
Heatmap



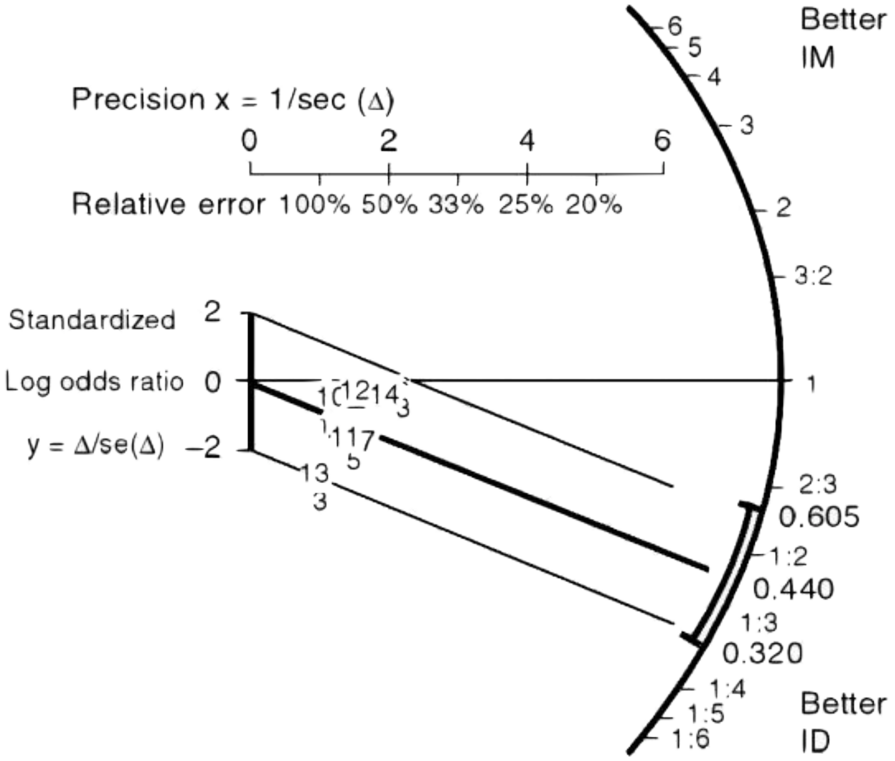
.. or just a table?

	Academic	Industrial
Empirically based	7.0	41.0
Empirically evaluated	8.5	2.5

L’Abbé plot



Galbraith (radial) plots



Economic assessment

- Full economic evaluation studies
  - Comparative analysis of alternative courses of action in terms of both
    - costs (resource use) and
    - consequences (outcomes, effects)
- Partial economic evaluation studies
  - No explicit comparisons between alternative interventions
- Single effectiveness studies
  - Limited information relating to the description, measurement or valuation of resource use associated with interventions

Full economic evaluation

- Cost-effectiveness analysis (CEA):
  - Effects of interventions are measured in identical units of outcome (e.g. defects).
  - Alternative interventions are compared in terms of ‘cost per unit of effect’.
- Cost-utility analysis (CUA):
  - Effects are expressed in utilities, alternative interventions may produce different levels of effect in terms of both quantity and quality of life (or different effects),.
  - Utilities are measures which comprise both length of life and subjective levels of well-being. The best known utility measure is the quality-adjusted life year, or QALY. Alternative interventions are compared in terms of cost per unit of utility gained (e.g. cost per QALY).
- Cost-benefit analysis (CBA):
  - Both resource inputs and effects of alternative interventions are expressed in monetary units,

SYSTEMATIC LITERATURE MAPPING

Systematic Literature Mapping

- A systematic mapping study provides a structure of the type of research reports and results that have been published by categorizing them.
  - It often gives a visual summary, the map, of its results.
  - It requires less effort while providing a more coarse-grained overview.

SLM vs. SLR

- Goals
- Process
- Breadth and depth
- Topic taxonomy
- Approach taxonomy
- Validity
- Industrial relevance

SLM vs SLR: Goals

- Research Questions
  - SLR: few specific
  - SLM: several broad
- Goals
  - SLR focus on identifying best practices based on empirical evidence
  - SLM focus on classification, conducting thematic analysis and identifying publication fora, spot research gaps

SLM vs. SLR: Process

- SLM
  - Articles are not evaluated regarding their quality as the main goal is not to establish the state of evidence
  - Thematic analysis, it helps to see which categories are well covered in terms of number of publications
    - Very different w.r.t. meta-analysis

SLM vs. SLR: breadth and depth

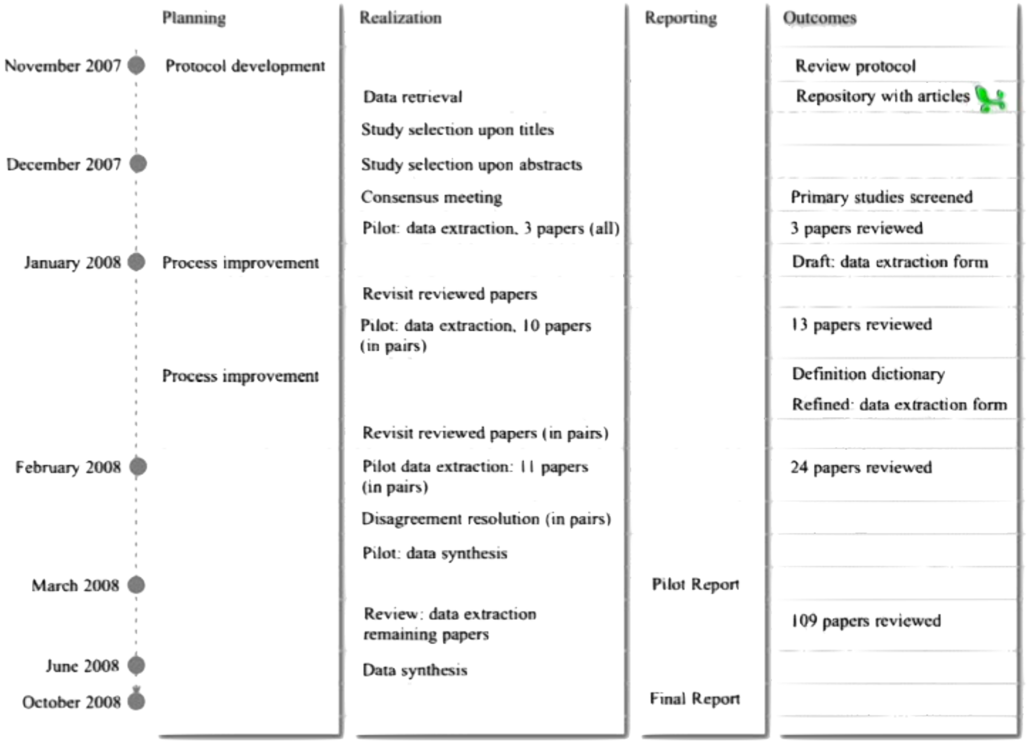
- SLM:
  - More articles can be considered as they don't have to be evaluated in detail.
  - Cover a larger field
    - Looser search strings

LESSONS LEARNED

Lessons Learned

- The poor quality of search engines available (precision, available fields)
- Researchers should familiarize themselves with how each search engine handles search terms.
- To avoid redundant searches, researchers should first plan which terms will be applied to which search engines and once completed, the results and timestamp are recorded.
- Due to the apparent fragility of some search engines a patient and opportunistic approach must be adopted.
- The variable quality of the abstracts available for Software Engineering papers
- More lessons learned in Brereton et al.(2007)

Illustrate timeline



	Planning	Realization	Reporting	Outcomes
November 2007	Protocol development			Review protocol
		Data retrieval		Repository with articles
		Study selection upon titles		
December 2007		Study selection upon abstracts		
		Consensus meeting		Primary studies screened
		Pilot: data extraction, 3 papers (all)		3 papers reviewed
January 2008	Process improvement			Draft: data extraction form
		Revisit reviewed papers		
		Pilot: data extraction, 10 papers (in pairs)		13 papers reviewed
	Process improvement			Definition dictionary
				Refined: data extraction form
		Revisit reviewed papers (in pairs)		
February 2008		Pilot data extraction: 11 papers (in pairs)		24 papers reviewed
		Disagreement resolution (in pairs)		
		Pilot: data synthesis		
March 2008			Pilot Report	
		Review: data extraction remaining papers		109 papers reviewed
June 2008		Data synthesis		
October 2008			Final Report	

Meta-evidence

- The software engineering research community is starting to adopt
- SLRs consistently as a research method.
  - number of SLRs is increasing.
  - number of researchers and organizations performing them is increasing.
- The integration of the results of the primary studies was poorly conducted by many SLRs.

Source (da Silva et al., 2011)

Meta-evidence

- There was very little consistency in the way the SLRs are organized.
- Many SLRs omitted essential data, including important parts of the review protocol.
- The majority of the SLRs:
  - not evaluate the quality of primary studies.
  - to provide guidelines for practitioners, thus decreasing their potential impact on software engineering practice.

Source (da Silva et al., 2011)

Summary

- Many of the steps in a systematic review assume that it will be undertaken by a large group of researchers.
- In the case of a PhD student, the most important steps to undertaken are:
    - Developing a protocol
    - Defining the research question
    - Specifying what will be done to address the problem of a single research applying inclusion/exclusion criteria and undertaking all the data extraction
    - Defining the search strategy
    - Defining the data to be extracted from each primary study including quality data
    - Maintaining list of included and excluded studies
    - Using the data synthesis guidelines
    - Using the reporting guidelines

Acknowledgment

- I wish to thank Prof. Marcela Genero for sharing with me her materials on SLR.
- The present slides are partly based on her work.

Bibliography

- Kitchenham, B. (2004). Procedures for Performing Systematic Reviews. Joint Technical Report TR/SE-0401.
- Kitchenham, B. and Charters, S. (2007). Guidelines for performing Systematic Literature Reviews in Software Engineering. Version 2.3 EBSE-2007-01.
- Brereton et al., (2007). Lessons from applying the systematic literature review process within the software engineering domain. Journal of Systems and Software, 80, 571-583.

Bibliography

- D. S. Cruzes, T. Dybå, *Recommended steps for thematic synthesis in software engineering*, in: 2011 International Symposium on Empirical Software Engineering and Measurement (ESEM), 2011, pp.275-284
- C. Wohlin, *Guidelines for snowballing in systematic literature studies and a replication in software engineering*, in: Proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering (EASE), 2014, 10pp

Bibliography

- Kitchenham, B., Brereton, P., Budgen,D., Turner,M., Bailey, J., Linkman, S. (2009). Systematic literature reviews in software engineering – A systematic literature review. Information and Software Technology 51 7-15.
- Kitchenham, B. e tal.(2010). Literature reviews in software engineering – a tertiary study, Information and Software Technology 52 (8) 792–805.
- Fabio Q.B. da Silva, André L.M. Santos, Sérgio Soares, A. César C. França, Cleviton V.F. (2011). Six Years of Systematic Literature Reviews in Software Engineering: An Updated Tertiary Study Information and Software Technology 53(9) 899–913.

Example

- Tore Dybå, Erik Arisholm, Dag I.K. Sjøberg, Jo E. Hannay, and Forrest Shull. “Are Two Heads Better than One? On the Effectiveness of Pair Programming”, *IEEE Software*, vol. 24, no. 6, 2007.
- Tore Dybå, Erik Arisholm, Dag I.K. Sjøberg, Jo E. Hannay, and Forrest Shull, "Method: How We Selected and Analyzed the Studies," *IEEE Software*, vol. 24, no. 6, 2007,
  - http://www2.computer.org/cms/Computer.org/dl/mags/so/2007/06/extras/mso2007060012x1.html