LTAT.05.008: Software Analytics

Systematic literature reviews, mapping studies, and questionnaire-based surveys

Dietmar Pfahl
email: dietmar.pfahl@ut.ee

Spring 2018
Prepare for Quiz 1

Read article:

Quiz 1: Software Analytics Introduction

(10 min -- 2 marks)
Data Collection & Research Methods in Data Science

• Survey
  – Questionnaire-based (primary study)
  – Literature-based (secondary / tertiary study)

• Case Study
  – Exploratory
  – Descriptive
  – Confirmatory/Explanatory
  – Improving

• Experiment
  – Controlled Experiment
  – Quasi-Experiment

• Many other...
  – Action Research
  – Ethnography
  – Longitudinal Studies
  – Design Science
SLR – Process

Phase 1: Plan Review

1. Specify Research Questions
2. Develop Review Protocol
3. Validate Review Protocol

Phase 2: Conduct Review

4. Identify Relevant Research
5. Select Primary Studies
6. Assess Study Quality
7. Extract Required Data
8. Synthesise Data

Phase 3: Document Review

9. Write Review Report
10. Validate Report
SLR – Process:
1. Specify Research Question

Good RQs are the cornerstone of the study!

- 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
SLR – Process:
1. Specify Research Question

- Taxonomy of types of RQs -> Lecture 1

Typically RQs in Software Engineering aim at:
- 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
SLR – Process:  
1. Specify Research Question

Good RQs are:

• Meaningful to SE practitioners as well as SE researchers
• Lead either to
  – changes in current software engineering practice
  or to
  – increased confidence in the value of current practice
• Identify discrepancies between commonly held beliefs and reality
SLR – Process:
2. Develop Review Protocol

• Background
• The research questions
• The strategy that will be used to search for primary studies including search terms and resources to be searched
• Study selection criteria
• Study selection procedures
• Study quality assessment checklists and procedures
• Data extraction strategy
• Synthesis of the extracted data
• Dissemination strategy
• Project timetable

An example protocol for a tertiary study can be found on the course wiki.
SLS – Process:
3. Validate Review Protocol

• If enough effort (funding) is available:
  – a group of independent experts should be asked to review the protocol
  – same experts could later be asked to review the final report
SLS – Process:
3. Validate Review Protocol

• Typical checks include:
  – Are the research questions clear/meaningful and aligned to the objectives of the SLR?
  – Are the search strings appropriately derived from the research questions?
  – Are proper quality criteria defined to assess the selected literature?
  – Will the data to be extracted properly address the research question(s)?
  – Is the data analysis procedure appropriate for the data to be extracted?
  – Is the data analysis procedure appropriate to answer the research questions?
SLS – Process:
4. Identify Relevant Research

- Manually assemble known relevant literature (primary studies)
- Identify data sources for automatic extraction
- Define search strings for automatic extraction
- Check precision & sensitivity with help of quasi gold standard

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

**4. Identify Relevant Research**

- Manually assemble known relevant literature (primary studies)
- Identify data sources for automatic extraction
- Define search strings for automatic extraction
- Check precision & sensitivity with help of quasi gold standard

If relevant studies are known, then forward and backward snowballing can be used to find more relevant studies.

Primary Study A

- References
  - Ref1
  - Ref2
  - ...

Cited in:

- Study B
- Study C
- ...

(Citation in text)

Cited in:

- Study B
- Study C
- ...

(Citation in text)
SLS – Process: 4. Identify Relevant Research

- Manually assemble known relevant literature (primary studies)
- Identify data sources for automatic extraction
- Define search strings for automatic extraction
- Check precision & sensitivity with help of quasi gold standard
SLS – Process: 4. Identify Relevant Research

• Manually assemble known relevant literature (primary studies)

• Identify data sources for automatic extraction

• Define search strings for automatic extraction

• Check precision & sensitivity with help of quasi gold standard

Software Engineering Digital Libraries

• IEEEExplore
• ACM Digital library
• SpringerLink
• Google scholar (scholar.google.com)
• Citeseer library (citeseer.ist.psu.edu)
• Inspec (www.iee.org/Publish/INSPEC/)
• ScienceDirect (www.sciencedirect.com)
• EI Compendex (www.engineeringvillage2.org/Controller/Servlet/AthensService)
SLS – Process:
4. Identify Relevant Research

- Manually assemble known relevant literature (primary studies)
- Identify data sources for automatic extraction
- Define search strings for automatic extraction
- Check precision & sensitivity with help of quasi gold standard

Dedicated SE sources:
- Empirical Software Engineering (J)
- Information and Software Technology (J)
- Journal of Software: Evolution and Process (J)
- Journal of Systems and Software (J)
- International Symposium on Empirical Software Engineering and Measurement (C)
- International Conference on Software Engineering (C)
- Evaluation and Assessment in Software Engineering (C)
SLS – Process:
4. Identify Relevant Research

- Manually assemble known relevant literature (primary studies)
- Identify data sources for automatic extraction
- Define search strings for automatic extraction
- Check precision & sensitivity with help of quasi gold standard

A general approach is to break down the question into individual facets, i.e. population, intervention, comparison, outcomes, context (PICOC). Then draw up a list of synonyms, abbreviations, and alternative spellings; other terms can be obtained by considering subject headings used in journals and data bases. Search strings can then be constructed using Boolean ANDs and ORs.
SLR – Process:
4. Identify Relevant Research

RQs can be structured using the PICOC criteria:

- Population
- Intervention
- Comparison
- Outcomes
- Context
SLR – Process: 4. Identify Relevant Research

RQs can be structured using the PICOC criteria:

- **Population**
- **Intervention**
- **Comparison**
- **Outcomes**
- **Context**

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.
SLR – Process:
4. Identify Relevant Research

RQs can be structured using the PICOC criteria:

- **Population**
- **Intervention**
- **Comparison**
- **Outcomes**
- **Context**

The intervention is the software methodology/tool/technology/procedure that addresses a specific issue.

Examples:
- technologies to perform specific tasks such as requirements specification, system testing, or software cost estimation
SLR – Process:
4. Identify Relevant Research

RQs can be structured using the PICOC criteria:

- **Population**
- **Intervention**
- **Comparison**
- **Outcomes**
- **Context**

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; “not using the intervention” is inadequate as a description of the control treatment

If training is required, beware of confounding factors (i.e., training effect versus effect of the technique)
SLR – Process:
4. Identify Relevant Research

RQs can be structured using the PICOC criteria:

- **Population**
- **Intervention**
- **Comparison**
- **Outcomes**
- **Context**

Outcomes should relate to factors of importance to practitioners such as improved reliability, reduced production costs, reduced time to market, …

- All relevant outcomes must be specified; 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 problem is the use of surrogate measures (e.g., defects found during system testing as a surrogate for quality, or coupling measures for design quality; using surrogate measures may be misleading and conclusions based on such studies may be less robust.)
SLR – Process: 4. Identify Relevant Research

RQs can be structured using the PICOC criteria:

- **Population**
- **Intervention**
- **Comparison**
- **Outcomes**
- **Context**

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).

Note: Many software experiments take place in academia using student participants and small scale tasks; such experiments might not be representative of what might occur with practitioners working in industry.
SLR – Process:
4. Identify Relevant Research

Exercise
What are the PICOC criteria in the following RQs?

- RQ 1: What evidence is there that cross-company estimation models are not significantly different from within-company estimation models for predicting effort for software/Web projects?
- RQ 2: What characteristics of the study data sets and the data analysis methods used in the study affect the outcome of within- and cross-company effort estimation accuracy studies?

Note: “study” refers to the primary study reported in the literature that is reviewed.
SLS – Process: 4. Identify Relevant Research

- Manually assemble known relevant literature (primary studies)
- Identify data sources for automatic extraction
- Define search strings for automatic extraction
- Check precision & sensitivity with help of quasi gold standard

Example:

- **Population:** software OR application OR product OR Web OR WWW OR Internet OR World-Wide Web OR project OR development
- **Intervention:** (cross company OR cross organisation OR cross organization OR multiple-organizational OR multiple-organisational model) AND (modeling OR modelling effort OR cost OR resource estimation OR prediction OR assessment)
- **Contrast:** within-organisation OR within-organization OR within-organizational OR within-organisational OR single company OR single organisation
- **Outcome:** Accuracy OR Mean Magnitude Relative Error

The search strings were constructed by linking the four OR lists using the Boolean AND.

RQ 1: What evidence is there that cross-company estimation models are not significantly different from within-company estimation models for predicting effort for software/Web projects?

RQ 2: What characteristics of the study data sets and the data analysis methods used in the study affect the outcome of within- and cross-company effort estimation accuracy studies?
SLS – Process: 4. Identify Relevant Research

• Manually assemble known relevant literature (primary studies)

• Identify data sources for automatic extraction

• Define search strings for automatic extraction

• Check precision & sensitivity with help of quasi gold standard

Example string for ACM DL:

‘+’ corresponds to AND; ‘blank’ corresponds to OR;
“…” corresponds to a phrase

+(software application product Web WWW Internet “World-Wide Web” project development)
+(“cross company” “cross organisation” “cross organization”
“multiple-organizational” “multiple-organisational model”)
+(modeling “modelling effort” cost “resource estimation”
prediction assessment)
+(“within-organisation” “within-organization” “within-
organizational” “within-organisational” “single company” “single
organisation”)
+(Accuracy “Mean Magnitude Relative Error”)

Effort estimation: how valuable is it for a web company to use a cross-company data set, compared to using its own single-company data set?

Emilia Mendes, Sergio Di Martino, Filomena Ferrucci, Carmine Gravino


Publisher: ACM

Bibliometrics: Citation Count: 7

Downloads (6 Weeks): 0, Downloads (12 Months): 6, Downloads (Overall): 465

Full text available: PDF

Previous studies comparing the prediction accuracy of effort models built using Web cross- and single-company data sets have been inconclusive, and as such replicated studies are necessary to determine under what circumstances a company can place reliance on a cross-company effort model. This paper therefore replicates a
SLS – Process: 4. Identify Relevant Research

- Manually assemble known relevant literature (primary studies)
- Identify data sources for automatic extraction
- Define search strings for automatic extraction
- Check precision & sensitivity with help of quasi gold standard
SLS – Process: 4. Identify Relevant Research

- Manually assemble known relevant literature (primary studies)
- Identify data sources for automatic extraction
- Define search strings for automatic extraction
- Check precision & sensitivity with help of quasi gold standard

High “quasi-sensitivity”: retrieves most of the studies in the “gold standard” but also many false positives.

High “quasi-precision”: retrieves only a few irrelevant primary studies but may miss many elements of the “gold standard”
SLS – Process:
4. Identify Relevant Research

- Manually assemble known relevant literature (primary studies)
- Identify data sources for automatic extraction
- Define search strings for automatic extraction
- Check precision & sensitivity with help of quasi gold standard

Sensitivity = \( \frac{9}{10} = 90\% \)

Precision = \( \frac{9}{20} = 45\% \)
SLS – Process:
4. Identify Relevant Research

- Manually assemble known relevant literature (primary studies)
- Identify data sources for automatic extraction
- Define search strings for automatic extraction
- Check precision & sensitivity with help of quasi gold standard

Sensitivity = 4/10 = 40%
Precision = 4/5 = 80%
SLS – Process:
4. Identify Relevant Research

• Manually assemble known relevant literature (primary studies)

• Identify data sources for automatic extraction

• Define search strings for automatic extraction

• Check precision & sensitivity with help of quasi gold standard
SLS – Process:
5. Select Primary Studies

• Study selection criteria are intended to identify those primary studies that provide direct evidence about the research question
  – selection criteria should be decided during the protocol definition
  – they should be piloted to ensure that they can be reliably interpreted and that they classify studies correctly
  – they may be refined during the search process
SLS – Process: 5. Select Primary Studies

- Inclusion and exclusion criteria should be based on the research question

Example:
- 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.

RQ 1: What evidence is there that cross-company estimation models are not significantly different from within-company estimation models for predicting effort for software/Web projects?

RQ 2: What characteristics of the study data sets and the data analysis methods used in the study affect the outcome of within- and cross-company effort estimation accuracy studies?
SLS – Process:
5. Select Primary Studies

• Additional Inclusion/Exclusion criteria might be based on
  – Language,
  – publication venue,
  – authors,
  – setting,
  – participants or subjects,
  – research design,
  – sampling method,
  – date of publication,
  – availability of full-text,
  – ...

If more than one researcher makes inclusion/exclusion decisions, then calculate inter-rater agreement (Cohen’s Kappa)
RQ: What requirements elicitation techniques are most effective?

SLS – Process:
5. Select Primary Studies

• Example

<table>
<thead>
<tr>
<th>SCOPUS</th>
<th>IEEEXPLORE</th>
<th>ACM Digital Library</th>
<th>Google</th>
</tr>
</thead>
<tbody>
<tr>
<td>(TITLE-ABS-KEY(&quot;elicitation OR &quot;requirements gathering&quot; OR &quot;requirements acquisition&quot;) AND TITLE-ABS-KEY(capture OR empirical OR experiment OR study OR review OR evaluation)) AND SUBJAREA engin OR soc OR psych OR ebuma</td>
<td>(elicitation &lt;OR&gt; &quot;requirements gathering&quot; &lt;OR&gt; &quot;requirements acquisition&quot;) &lt;AND&gt; (capture &lt;OR&gt; empirical &lt;OR&gt; experiment &lt;OR&gt; study &lt;OR&gt; review &lt;OR&gt; evaluation) AND elicitation</td>
<td>(capture OR empirical OR experiment OR study OR review OR evaluation) AND &quot;requirements gathering&quot;</td>
<td>(capture OR empirical OR experiment OR study OR review OR evaluation) AND &quot;requirements acquisition&quot;</td>
</tr>
</tbody>
</table>

737 (pubs. reported by search)
474
37 (candidate pubs.)

197
176
89
92

Unspecified

The references are reviewed recursively. The review ends when (1) the references belong to a completely different field of knowledge or (2) they are too dated. Both criteria tend to apply simultaneously.

74 candidate publications

15 selected publications

52 publications of no interest
7 unavailable publications

460 publications of no interest
19 unavailable publications

Backward Snowballing
SLS – Process:
6. Assess Quality of Primary Studies

• Even more detailed inclusion/exclusion criteria
  – Only include studies that have a minimum quality

• Investigate whether quality differences provide an explanation for differences in results of primary studies

• 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
SLS – Process:
6. Assess Quality of Primary Studies

• Quality checklists aim at spotting factors that could bias primary study results

<table>
<thead>
<tr>
<th>Type</th>
<th>Synonyms</th>
<th>Definition</th>
<th>Protection mechanism</th>
</tr>
</thead>
<tbody>
<tr>
<td>Selection bias</td>
<td>Allocation</td>
<td>Systematic difference between comparison groups with respect to treatment</td>
<td>Randomisation of a large number of subjects with concealment of the allocation method (e.g. allocation by computer program not experimenter choice)</td>
</tr>
<tr>
<td>Performance bias</td>
<td></td>
<td>Systematic difference is the conduct of comparison groups apart from the treatment being evaluated.</td>
<td>Replication of the studies using different experimenters. Use of experimenters with no personal interest in either treatment.</td>
</tr>
<tr>
<td>Measurement bias</td>
<td>Detection Bias</td>
<td>Systematic difference between the groups in how outcomes are ascertained.</td>
<td>Blinding outcome assessors to the treatments is sometimes possible.</td>
</tr>
<tr>
<td>Attrition bias</td>
<td>Exclusion bias</td>
<td>Systematic differences between comparison groups in terms of withdrawals or exclusions of participants from the study sample.</td>
<td>Reporting of the reasons for all withdrawals. Sensitivity analysis including all excluded participants.</td>
</tr>
</tbody>
</table>
SLS – Process:
6. Assess Quality of Primary Studies

Example of tailored guidelines applied to primary studies:

1. Appropriateness of data analysis process
2. Sensitivity or residual analysis done
3. Accuracy statistics based on the raw data scale
4. Appropriateness of study comparison method
5. The size of the within-company data set, measured according to the criteria presented below
6. Reporting quality

RQ 1: What evidence is there that cross-company estimation models are not significantly different from within-company estimation models for predicting effort for software/Web projects?

RQ 2: What characteristics of the study data sets and the data analysis methods used in the study affect the outcome of within- and cross-company effort estimation accuracy studies?
SLS – Process:
6. Assess Quality of Primary Studies

Example of tailored guidelines:

1. Appropriateness of data analysis process:
   - 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?

RQ 1: What evidence is there that cross-company estimation models are not significantly different from within-company estimation models for predicting effort for software/Web projects?

RQ 2: What characteristics of the study data sets and the data analysis methods used in the study affect the outcome of within- and cross-company effort estimation accuracy studies?
SLS – Process: 6. Assess Quality of Primary Studies

Example of tailored guidelines:

2. Sensitivity or residual analysis done
   – 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?

RQ 1: What evidence is there that cross-company estimation models are not significantly different from within-company estimation models for predicting effort for software/Web projects?

RQ 2: What characteristics of the study data sets and the data analysis methods used in the study affect the outcome of within- and cross-company effort estimation accuracy studies?
SLS – Process:
6. Assess Quality of Primary Studies

Example of tailored guidelines:

3. Accuracy statistics based on the raw data scale

E.g.: If data is on ordinal scale, cannot calculate mean error statistics

RQ 1: What evidence is there that cross-company estimation models are not significantly different from within-company estimation models for predicting effort for software/Web projects?

RQ 2: What characteristics of the study data sets and the data analysis methods used in the study affect the outcome of within- and cross-company effort estimation accuracy studies?
SLS – Process:
6. Assess Quality of Primary Studies

Example of tailored guidelines:

4. Appropriateness of study comparison method
   - Was the single company selected at random (not selected for convenience) from several different companies?
   - Was the comparison based on
     - cross-validation (1)
     - an independent hold out sample (0.5)
     - random subsets (0.33)
     - leave-one-out (0.17)
     - no hold out (0)

RQ 1: What evidence is there that cross-company estimation models are not significantly different from within-company estimation models for predicting effort for software/Web projects?
RQ 2: What characteristics of the study data sets and the data analysis methods used in the study affect the outcome of within- and cross-company effort estimation accuracy studies?
SLS – Process:
6. Assess Quality of Primary Studies

Example of tailored guidelines:

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)

RQ 1: What evidence is there that cross-company estimation models are not significantly different from within-company estimation models for predicting effort for software/Web projects?

RQ 2: What characteristics of the study data sets and the data analysis methods used in the study affect the outcome of within- and cross-company effort estimation accuracy studies?
SLS – Process:
6. Assess Quality of Primary Studies

Example of tailored guidelines:

6. Reporting quality:
   – Is it clear what projects were used to construct each model?
   – Is it clear how accuracy was measured?
   – Is it clear what cross-validation method was used?
   – Were all model construction methods fully defined (tools and methods used)?
   – ...

RQ 1: What evidence is there that cross-company estimation models are not significantly different from within-company estimation models for predicting effort for software/Web projects?

RQ 2: What characteristics of the study data sets and the data analysis methods used in the study affect the outcome of within- and cross-company effort estimation accuracy studies?
SLS – Process:
6. Assess Quality of Primary Studies

Elements of guidelines for different types of primary studies:

<table>
<thead>
<tr>
<th>Question</th>
<th>Quantitative Empirical Studies (no specific type)</th>
<th>Correlation (observational studies)</th>
<th>Surveys</th>
<th>Experiments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Design</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Are the aims clearly stated?</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Was the study designed with these questions in mind?</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Do the study measures allow the questions to be answered?</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>What population was being studied?</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Who was included?</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Who was excluded?</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>How was the sample obtained (e.g., postal, interview, web-based)?</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Is the survey method likely to have introduced significant bias?</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Is the sample representative of the population to which the results will generalise?</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Were treatments randomly allocated?</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Is there a comparison or control group?</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>If there is a comparison or control group, are participants similar to the treatment group participants in terms of variables that may affect study outcomes?</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Was the sample size justified?</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>If the study involves assessment of a technology, is the technology clearly defined?</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Could the choice of subjects influence the size of the treatment effect?</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Could lack of blinding introduce bias?</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Are the variables used in the study adequately measured (i.e., are the variables likely to be valid and reliable)?</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Are the measures used in the study fully defined?</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Are the measures used in the study the most relevant ones for answering the research questions?</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Is the scope (size and length) of the study sufficient to allow for changes in the outcomes of interest to be identified?</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
</tbody>
</table>
SLS – Process:
6. Assess Quality of Primary Studies

Elements of guidelines for different types of primary studies:

<table>
<thead>
<tr>
<th>Question</th>
<th>Quantitative Empirical Studies (no specific type)</th>
<th>Correlation (observational studies)</th>
<th>Surveys</th>
<th>Experiments</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Conduct</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Did untoward events occur during the study?</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Was outcome assessment blind to treatment group?</td>
<td>X</td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>Are the data collection methods adequately described?</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>If two groups are being compared, were they treated similarly within the study?</td>
<td>X</td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>If the study involves participants over time, what proportion of people who enrolled at the beginning dropped out?</td>
<td></td>
<td></td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>How was the randomisation carried out?</td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td><strong>Analysis</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>What was the response rate?</td>
<td></td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Was the denominator (i.e., the population size) reported?</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Do the researchers explain the data types (continuous, ordinal, categorical)?</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Are the study participants or observational units adequately described? (For example, SE experience, type (student, practitioner, consultant), nationality, task experience and other relevant variables)</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Were the basic data adequately described?</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Have “drop outs” introduced bias?</td>
<td>X</td>
<td></td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Are reasons given for refusal to participate?</td>
<td>X</td>
<td></td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Are the statistical methods described?</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Is the statistical program used to analyse the data referenced?</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Are the statistical methods justified?</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Is the purpose of the analysis clear?</td>
<td>X</td>
<td></td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Are scoring systems described?</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Are potential confounders adequately controlled for in the analysis?</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Do the numbers add up across different tables and subgroups?</td>
<td>X</td>
<td></td>
<td>X</td>
<td>X</td>
</tr>
</tbody>
</table>
### SLS – Process:

#### 6. Assess Quality of Primary Studies

Elements of guidelines for different types of primary studies:

<table>
<thead>
<tr>
<th>Question</th>
<th>Quantitative Empirical Studies (no specific type)</th>
<th>Correlation (observational studies)</th>
<th>Surveys</th>
<th>Experiments</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Conduct</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>If different groups were different at the start of the study or treated differently during the study, was any attempt made to control for these differences, either statistically or by matching?</td>
<td>X</td>
<td></td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>If yes, was it successful?</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Was statistical significance assessed?</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>If statistical tests are used to determine differences, is the actual p value given?</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>If the study is concerned with differences among groups, are confidence limits given describing the magnitude of any observed differences?</td>
<td>X</td>
<td></td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Is there evidence of multiple statistical testing or large numbers of post hoc analyses?</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>How could selection bias arise?</td>
<td>X</td>
<td></td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Were side-effects reported?</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Conclusions</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Are all study questions answered?</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>What do the main findings mean?</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Are negative findings presented?</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>If statistical tests are used to determine differences, is practical significance discussed?</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>If drop outs differ from participants, are limitations to the results discussed?</td>
<td>X</td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>How are null findings interpreted? (I.e. has the possibility that the sample size is too small been considered?)</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Are important effects overlooked?</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>How do results compare with previous reports?</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>How do the results add to the literature?</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>What implications does the report have for practice?</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Do the researchers explain the consequences of any problems with the validity/reliability of their measures?</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
</tbody>
</table>
SLS – Process: 7. Extract Required Data

• Design data extraction forms to accurately record the information researchers obtain from the primary studies
  – Linked to the research questions
  – Must address the quality criteria

• Usually, electronic forms (spreadsheets) are used
  – Advanced tool support provides hyperlinks from spreadsheet to primary studies
SLS – Process:

7. Extract Required Data

• Example:

RQ 1: What evidence is there that cross-company estimation models are not significantly different from within-company estimation models for predicting effort for software/Web projects?

RQ 2: What characteristics of the study data sets and the data analysis methods used in the study affect the outcome of within- and cross-company effort estimation accuracy studies?
### SLS – Process

#### 7. Extract

**Example:**

RQ 1: What evidence is there that cross-company estimation models are not significantly different from within-company estimation models for predicting effort for software/Web projects?

RQ 2: What characteristics of the study data sets and the data analysis methods used in the study affect the outcome of within- and cross-company effort estimation accuracy studies?

<table>
<thead>
<tr>
<th>Cross-company model</th>
<th>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.</th>
</tr>
</thead>
<tbody>
<tr>
<td>What technique(s) was used to construct the cross-company model?</td>
<td>In all cases, accuracy assessment was based on the logarithmic models not the additive models.</td>
</tr>
<tr>
<td>If several techniques were used which was most accurate?</td>
<td>Not clear whether the variables were transformed or the GLM was used to construct a log-linear model</td>
</tr>
<tr>
<td>What transformations if any were used?</td>
<td>It can be assumed that linear models did not work well.</td>
</tr>
<tr>
<td>What variables were included in the cross-company model?</td>
<td>KLOC, Language subset, Category subset, RELY</td>
</tr>
<tr>
<td>What cross-validation method was used?</td>
<td>A hold-out sample of 9 projects from the single company was used to assess estimate accuracy</td>
</tr>
<tr>
<td>Was the cross-company model compared to a baseline to check if it was better than chance?</td>
<td>Yes</td>
</tr>
<tr>
<td>What was/were the measure(s) used as benchmark?</td>
<td>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.)</td>
</tr>
</tbody>
</table>
SLS – Process:
7. Extract

• Example:

| RQ 1: What evidence is there that cross-company estimation models are not significantly different from within-company estimation models for predicting effort for software/Web projects? |
| RQ 2: What characteristics of the study data sets and the data analysis methods used in the study affect the outcome of within- and cross-company effort estimation accuracy studies? |

<table>
<thead>
<tr>
<th>Within-company model</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>What technique(s) was used to construct the within-company model?</td>
<td>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.</td>
</tr>
<tr>
<td>If several techniques were used which was most accurate?</td>
<td>In all cases, accuracy assessment was based on the logarithmic models not the additive models.</td>
</tr>
<tr>
<td>What transformations if any were used?</td>
<td>Not clear whether the variables were transformed or the GLM was used to construct a log-linear model.</td>
</tr>
<tr>
<td>What variables were included in the within-company model?</td>
<td>KLOC, Language subset, Year</td>
</tr>
<tr>
<td>What cross-validation method was used</td>
<td>A hold-out sample of 9 projects from the single company was used to assess estimate accuracy</td>
</tr>
</tbody>
</table>
SLS – Process

7. Extract

Example:

RQ 1: What evidence is there that cross-company estimation models are not significantly different from within-company estimation models for predicting effort for software/Web projects?

RQ 2: What characteristics of the study data sets and the data analysis methods used in the study affect the outcome of within- and cross-company effort estimation accuracy studies?
SLS – Process: 7. Extract Required Data

**Example:**

<table>
<thead>
<tr>
<th>Data Summary</th>
<th>With-company data summary for size and effort metrics</th>
<th>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</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data base summary (all projects) for size and effort metrics.</td>
<td>Effort: min: 7.8 MM</td>
<td>Not specified</td>
</tr>
<tr>
<td></td>
<td>Effort: max: 4361 MM</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Effort: mean: 284 MM</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Effort median: 93 MM</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Size min: 2000 KLOC</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Size max: 413000 KLOC</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Size mean: 51010 KLOC</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Size median: 22300 KLOC</td>
<td></td>
</tr>
</tbody>
</table>

RQ 1: What evidence is there that cross-company estimation models are not significantly different from within-company estimation models for predicting effort for software/Web projects? RQ 2: What characteristics of the study data sets and the data analysis methods used in the study affect the outcome of within- and cross-company effort estimation accuracy studies?
SLS – Process:
7. Extract Required Data

Remarks about extraction:
• If several primary studies are included in one publication, each must be treated separately
  – If it’s unclear to which study data belongs, one may contact the authors for clarification
• If data is incomplete one may contact the authors
• If there are several publications about the same study, analyse the most recent/complete version
SLS – Process: 8. Synthesise Data

- Collating and summarising the results of the included primary studies
  - Descriptive (non-quantitative/narrative) Synthesis
  - Quantitative Synthesis

- If a primary study contains both, qualitative and quantitative data:
  - Analyse separately
  - Then integrate and try to use the qualitative results to support the quantitative results
SLS – Process:
8. Synthesise Data

Descriptive (non-quantitative/narrative) 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
SLS – Process: 8. Synthesise Data

Descriptive (non-quantitative/narrative) 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
SLS – Process: 8. Synthesise Data

Descriptive (non-quantitative/narrative) Synthesis

Thematic Synthesis Process

- 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

<table>
<thead>
<tr>
<th>Step</th>
<th>Description</th>
<th>Data Input</th>
<th>Data Output</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Initial reading of data/text</td>
<td>Many pages of text</td>
<td>30-40 codes</td>
</tr>
<tr>
<td>2</td>
<td>Identify specific segments of text</td>
<td>Many segments of text</td>
<td>15-20 themes</td>
</tr>
<tr>
<td>3</td>
<td>Label the segments of text</td>
<td>30-40 codes</td>
<td>5-7 themes</td>
</tr>
<tr>
<td>4</td>
<td>Reduce overlap and translate codes into themes</td>
<td>15-20 themes</td>
<td>5-7 themes</td>
</tr>
<tr>
<td>5</td>
<td>Create a model of higher-order themes</td>
<td>5-7 themes</td>
<td>Model</td>
</tr>
<tr>
<td>Steps</td>
<td>Description</td>
<td>Checklist</td>
<td></td>
</tr>
<tr>
<td>----------------------------</td>
<td>------------------------------------------------------------------------------</td>
<td>-----------------------------------------------------------------------------------------------------------------------------------------</td>
<td></td>
</tr>
<tr>
<td><strong>Text</strong></td>
<td>Extract data from the primary studies, including bibliographical information,</td>
<td>1. Have all papers been read carefully to get immersed with the data?</td>
<td></td>
</tr>
<tr>
<td></td>
<td>aims, context, and results.</td>
<td>2. Have specific segments of text pertaining to the objectives of the synthesis been identified?</td>
<td></td>
</tr>
<tr>
<td><strong>Codes</strong></td>
<td>Identify and code interesting concepts, categories, findings, and results in</td>
<td>3. Have publication details, context descriptions, and findings been extracted from all papers?</td>
<td></td>
</tr>
<tr>
<td></td>
<td>a systematic fashion across the entire data set.</td>
<td>4. Have another researcher checked the extraction?</td>
<td></td>
</tr>
<tr>
<td><strong>Themes</strong></td>
<td>Translate codes into themes, subthemes, and higher order themes.</td>
<td>5. Have important segments of text like concepts, categories, findings, and results been labeled and coded?</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>6. Has coding been done across the entire data set on a level that is appropriate for the research questions?</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>7. Has a list of initial codes with definitions and frequencies been created and checked by another researcher?</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>8. Have consistency checks or inter-rater reliability checks been performed to establish the credibility of the coding?</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>9. Are there clear, evident connections between the text and the codes?</td>
<td></td>
</tr>
<tr>
<td><strong>Model</strong></td>
<td>Create a model of higher-order themes and create a model of higher-order</td>
<td>10. Have themes been created from a thorough, inclusive, and comprehensive review of the codes of all papers?</td>
<td></td>
</tr>
<tr>
<td></td>
<td>themes.</td>
<td>11. Has overlap between codes been reduced and the remaining codes been collated and translated into themes?</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>12. Have themes been checked against each other and back to the data of the original papers?</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>13. Are themes internally coherent, consistent, and distinctive?</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Assess the trustworthiness of the interpretations leading up to the thematic</td>
<td>14. Have themes been compared across studies, translated into each other, and interpreted into higher-order themes?</td>
<td></td>
</tr>
<tr>
<td></td>
<td>synthesis.</td>
<td>15. Have higher-order themes and relationships between themes been checked against the research questions of the synthesis?</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>16. Are there clear descriptions of the higher-order themes and the relationships between these themes?</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>17. Has a model been created to show the relationships between the higher-order themes?</td>
<td></td>
</tr>
</tbody>
</table>
SLS – Process: 8. Synthesise Data

Descriptive (non-quantitative/narrative) Synthesis

Thematic Synthesis Process

Three types of data should be extracted:

• Publication details
  – e.g., authors, year, title, source, abstract, aims

• Context descriptions
  – e.g., subjects, technologies, industry, settings

• Findings
  – e.g., results, behaviors, actions, phenomena, events, quotes
SLS – Process: 8. Synthesise Data

Descriptive (non-quantitative/narrative) Synthesis

Thematic Synthesis Process

Example of a data extraction template
SLS – Process: 8. Synthesise Data

Descriptive (non-quantitative) Synthesis

Thematic Synthesis Process

Coding
- Deductive (a priori): fixed list of codes
- Inductive (grounded theory): develop set of codes bottom-up
- Integrated

RQ1: What are the characteristics of Software Engineers?
RQ2: What (de)motivates Software Engineers to be more (less) productive?
RQ3: What are the external signs or outcomes of (de)motivated Software Engineers?
RQ4: What aspects of Software Engineering (de)motivate Software Engineers?
RQ5: What models of motivation exist in Software Engineering?

To support this observation, one simply has to take notice of the turnover rates. Upon conclusion of each appraisal period, company personnel turnover rates miraculously peak [2]. This indicates that engineers seek recognition based on objective criteria and not based on management intuition or personal preferences [2].

Motivator: “Recognition based on objective criteria”

Code ➔ “Recognition”
SLS – Process: 8. Synthesise Data

Descriptive (non-quantitative/narrative) Synthesis

Thematic Synthesis Process

Thematic Map
SLS – Process:
8. Synthesise Data

Descriptive (non-quantitative/narrative) Synthesis

Thematic Synthesis Process

Integrated tool support (prototype)

Google Sheet

Publication

Mind Map
SLS – Process:
8. Synthesise Data

Descriptive (non-quantitative) Synthesis

Thematic Synthesis Process

- RQ1: What are the characteristics of Software Engineers?
- RQ2: What (de)motivates Software Engineers to be more (less) productive?
- RQ3: What are the external signs or outcomes of (de)motivated Software Engineers?
- RQ4: What aspects of Software Engineering (de)motivate Software Engineers?
- RQ5: What models of motivation exist in Software Engineering?

Example: Model (hierarchical)
SLS – Process: 8. Synthesise Data

Quantitative Synthesis

• Report at least the following:
  – Sample size for each intervention
  – Estimated 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
SLS – Process: 8. Synthesise Data

Quantitative Synthesis

- Odds ratio for binary data
- Effect size for continuous data

SLS – Process:  
8. Synthesise Data

Remarks:

• Often difficult to combine results due to different conceptualizations (qualitative) and different measures (quantitative) used
  – E.g., different criteria for success/failure, different definition of ‘productivity’, …

• Try to avoid post-hoc analysis (i.e., fishing for results) by specifying already in the review protocol assumptions (hypotheses) on how you expect interventions impact the outcomes
SLS – Process:
9. Write Review Report

• Using the specified dissemination channel:
  – Journal, conference/workshop paper, technical report, etc.

• Typical structure:
  – See next 2 slides
**SLS – Process:**

**9. Write Review Report**

### Structure

<table>
<thead>
<tr>
<th>Section</th>
<th>Subsection</th>
<th>Scope</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Title</strong></td>
<td></td>
<td></td>
<td>The title should be short but informative. It should be based  on the question being asked. In journal papers, it should indicate that the study is a systematic review.</td>
</tr>
<tr>
<td><strong>Authorship</strong></td>
<td></td>
<td></td>
<td>When research is done collaboratively, criteria for determining both who should be credited as an author, and the order of author’s names should be defined in advance. The contribution of workers not credited as authors should be noted in the Acknowledgements section.</td>
</tr>
<tr>
<td><strong>Executive summary</strong></td>
<td><strong>Context</strong></td>
<td>The importance of the research questions addressed by the review.</td>
<td>A structured summary or abstract allows readers to assess quickly the relevance, quality and generality of a systematic review.</td>
</tr>
<tr>
<td></td>
<td><strong>Objectives</strong></td>
<td>The questions addressed by the systematic review.</td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>Methods</strong></td>
<td>Data Sources, Study selection, Quality Assessment and Data extraction.</td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>Results</strong></td>
<td>Main finding including any meta-analysis results and sensitivity analyses.</td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>Conclusions</strong></td>
<td>Implications for practice and future research.</td>
<td></td>
</tr>
<tr>
<td><strong>Background</strong></td>
<td></td>
<td>Justification of the need for the review.</td>
<td>Description of the software engineering technique being investigated and its potential importance.</td>
</tr>
<tr>
<td></td>
<td><strong>Review questions</strong></td>
<td>Each review question should be specified.</td>
<td>Identify primary and secondary review questions. Note this section may be included in the background section.</td>
</tr>
<tr>
<td><strong>Review Methods</strong></td>
<td><strong>Data sources and search strategy</strong></td>
<td></td>
<td>This should be based on the research protocol. Any changes to the original protocol should be reported.</td>
</tr>
<tr>
<td></td>
<td><strong>Study selection</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>Study quality assessment</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>Data extraction</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>Data synthesis</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Included and excluded studies</strong></td>
<td><strong>Inclusion and exclusion criteria.</strong></td>
<td>List of excluded studies with rationale for exclusion.</td>
<td>Study inclusion and exclusion criteria can sometimes best be represented as a flow diagram because studies will be excluded at different stages in the review for different reasons.</td>
</tr>
</tbody>
</table>

---

*LTAT.05.008 / Software Analytics / © Dietmar Pfahl 2018*
## SLS – Process:
### 9. Write Review Report

### Structure

<table>
<thead>
<tr>
<th>Section</th>
<th>Subsection</th>
<th>Scope</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Results</td>
<td>Findings</td>
<td>Description of primary studies. Results of any quantitative summaries</td>
<td>Non-quantitative summaries should be provided to summarise each of the</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Details of any meta-analysis.</td>
<td>studies and presented in tabular form.</td>
</tr>
<tr>
<td></td>
<td>Sensitivity</td>
<td>Analysis</td>
<td>Quantitative summary results should be presented in tables and graphs.</td>
</tr>
<tr>
<td></td>
<td>Analysis</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Discussion</td>
<td>Principal</td>
<td>Findings</td>
<td>These must correspond to the findings discussed in the results section.</td>
</tr>
<tr>
<td></td>
<td>findings</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Strengths</td>
<td>Strengths and weaknesses of the evidence included in the review.</td>
<td>A discussion of the validity of the evidence considering bias in the</td>
</tr>
<tr>
<td></td>
<td>and Weaknesses</td>
<td>Relation to other reviews, particularly considering any differences in</td>
<td>systematic review allows a reader to assess the reliance that may be placed</td>
</tr>
<tr>
<td></td>
<td></td>
<td>quality and results.</td>
<td>on the collected evidence.</td>
</tr>
<tr>
<td></td>
<td>Meaning of</td>
<td>Direction and magnitude of effect observed in summarised studies.</td>
<td>Make clear to what extent the results imply causality by discussing the</td>
</tr>
<tr>
<td></td>
<td>findings</td>
<td>Applicability (generalisability) of the findings.</td>
<td>level of evidence.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Discuss all benefits, adverse effects and risks.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Discuss variations in effects and their reasons (for example are the</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>treatment effects larger on larger projects).</td>
</tr>
<tr>
<td>Conclusions</td>
<td>Recommendations</td>
<td>Practical implications for software development.</td>
<td>What are the implications of the results for practitioners?</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Unanswered questions and implications for future research.</td>
<td></td>
</tr>
<tr>
<td>Acknowledgements*</td>
<td></td>
<td>All persons who contributed to the research but did not fulfil authorship</td>
<td></td>
</tr>
<tr>
<td>Conflict of Interest</td>
<td></td>
<td>Any secondary interest on the part of the researchers (e.g. a financial</td>
<td></td>
</tr>
<tr>
<td>References</td>
<td>Appendices</td>
<td>Appendices can be used to list studies included and excluded from the</td>
<td></td>
</tr>
<tr>
<td>and Appendices</td>
<td></td>
<td>study, to document search strategy details, and to list raw data from</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>the included studies.</td>
<td></td>
</tr>
</tbody>
</table>
SLS – Process:
10. Validate Report

• Before submission to a (formal) publication venue:
  – Expert review (could be the same who reviewed the protocol)

• After submission:
  – (double-blind) peer review by other researchers
Mapping Studies (MS)

- Difference to SLR
  - MS are different in terms of goals, breadth, validity issues and implications
  - Thus, MS and SLR should be used complementarily and require different methods (e.g., for analysis)
MS – Process:
1. Definition of Research Questions

• Primary goal:
  – to provide an overview of a research area
  – to identify the quantity and type of research and results available within it
  – often one wants to map the frequencies of publication over time to see trends

• Secondary goal:
  – to identify the forums in which research in the area of interest has been published
MS – Process:
1. Definition of Research Questions

Research Questions – 2 Examples:

<table>
<thead>
<tr>
<th>Object Oriented Design Map</th>
<th>Software Product Line Variability Map</th>
</tr>
</thead>
<tbody>
<tr>
<td>RQ1: Which journals include papers on software design?</td>
<td></td>
</tr>
<tr>
<td>RQ2: What are the most investigated object oriented design topics and how have these changed over time?</td>
<td></td>
</tr>
<tr>
<td>RQ3: What are the most frequently applied research methods, and in what study context?</td>
<td></td>
</tr>
<tr>
<td>RQ1: What areas in software product line variability are addressed and how many articles cover the different areas?</td>
<td></td>
</tr>
<tr>
<td>RQ2: What types of papers are published in the area and in particular what type of evaluation and novelty do they constitute?</td>
<td></td>
</tr>
</tbody>
</table>
MS – Process: 2. Conduct Search

Search Strings – 2 Examples:

• Object Oriented Design Map:
  – ("object oriented" AND "design" AND "empirical evidence") OR ("OO" AND "empirical" AND "design") OR ("software design" AND "OO" AND "experimental")

• Software Product Line Variability Map:
  – "software" AND ("product line" OR "product family" OR "system family") AND ("variability" OR "variation")
MS – Process:
3. Screening of Papers

Inclusion & Exclusion Criteria – 2 Examples:

<table>
<thead>
<tr>
<th>Object Oriented Design Map (Bailey et al. 2007)</th>
<th>Software Product Line Variability Map (Mujtaba et al. 2008)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Inclusion:</strong> books, papers, technical reports and grey literature describing empirical studies regarding object oriented software design. Where several papers reported the same study, only the most recent was included. Where several studies were reported in the same paper, each relevant study was treated separately. <strong>Exclusion:</strong> Studies that did not report empirical findings or literature that was only available in the form of abstracts or Powerpoint presentations.</td>
<td><strong>Inclusion:</strong> The abstract explicitly mentions variability or variation in the context of software product line engineering. From the abstract, the researcher is able to deduce that the focus of the paper contributes to product line variability research. <strong>Exclusion:</strong> The paper lies outside the software engineering domain. Variability and variation are not part of the contributions of the paper, the terms are only mentioned in the general introductory sentences of the abstract.</td>
</tr>
</tbody>
</table>
MS – Process:
4. Keywording using Abstracts

• Similar as for SLR

• Typically three facets:
  – Context facet – study-specific
  – Contribution facet – study-specific
  – Research type facet: pre-defined (Wieringa et al. 2006)
### 4. Keywording using Abstracts

#### Research Type Facet (Wieringa et al. 2006)

<table>
<thead>
<tr>
<th>Category</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Validation Research</td>
<td>Techniques investigated are novel and have not yet been implemented in practice. Techniques used are for example experiments, i.e., work done in the lab.</td>
</tr>
<tr>
<td>Evaluation Research</td>
<td>Techniques are implemented in practice and an evaluation of the technique is conducted. That means, it is shown how the technique is implemented in practice (solution implementation) and what are the consequences of the implementation in terms of benefits and drawbacks (implementation evaluation). This also includes to identify problems in industry.</td>
</tr>
<tr>
<td>Solution Proposal</td>
<td>A solution for a problem is proposed, the solution can be either novel or a significant extension of an existing technique. The potential benefits and the applicability of the solution is shown by a small example or a good line of argumentation.</td>
</tr>
<tr>
<td>Philosophical Papers</td>
<td>These papers sketch a new way of looking at existing things by structuring the field in form of a taxonomy or conceptual framework.</td>
</tr>
<tr>
<td>Opinion Papers</td>
<td>These papers express the personal opinion of somebody whether a certain technique is good or bad, or how things should been done. They do not rely on related work and research methodologies.</td>
</tr>
<tr>
<td>Experience Papers</td>
<td>Experience papers explain on what and how something has been done in practice. It has to be the personal experience of the author.</td>
</tr>
</tbody>
</table>
MS – Process:
4. Keywording using Abstracts

Building the Classification Scheme
MS – Process: 5. Data Extraction and Mapping Process
Questionnaire-based Surveys (QS)

- QS = Primary Study

- Process:
  - Define research goals/questions
  - Define target population and expected response rate
  - Define distribution channels and time frame
  - Define questionnaire (instruments)
  - Administer survey
  - Analyse data wrt. research questions
  - Report results
Questionnaire-based Surveys (QS)

When to use it?

• Either at start of research to get an understanding of the current situation …

• or at the end of a research phase to see the impact/acceptance/etc. of a new method/technique/tool

Issues:

• 'Superficial' --> no explanation / no causality --> not suitable for hypothesis testing

• 'Generalisability' of results depends on the choice of population and 'response rate', as well as validity and reliability of the data collection instrument
Questionnaire-based Surveys (QS)

Process:
• Setting specific, measurable objectives
• Planning and scheduling the survey
• Ensuring that appropriate resources are available
• Designing the survey
• Preparing the data collection instrument
• Validating the instrument
• Selecting participants
• Administering and scoring the instrument
• Analyzing the data
• Reporting the results
Questionnaire-based Surveys (QS)

Important issues:

• If population is too large to get response from all members, try to get a representative sample
  – Think about proper distribution channels
  – Check response-rate
  – Beware of self-selection bias

• Pilot the instrument
  – Beware of threats to construct validity
  – Construct validity = The ability of the instrument (or measure) to actually measure what you intend to measure
Questionnaire-based Surveys (QS)

Important issues:

- If population is too large to get response from all members, try to get a representative sample
  - Think about proper distribution channels
  - Check response-rate
  - Beware of self-selection bias
- Pilot the instrument
  - Beware of threats to construct validity
  - Construct validity = The ability of the instrument to actually measure what you intend to measure

How to improve response rate:
- Pose clear questions
- Make the questionnaire not too long
- Don’t ask intrusive questions
- Provide clear benefit for answering the questions
  (check this in pilot)
Subjective – Objective
Quantitative – Qualitative

<table>
<thead>
<tr>
<th>Qualitative (nominal, ordinal)</th>
<th>Subjective</th>
<th>Objective</th>
</tr>
</thead>
<tbody>
<tr>
<td>Quantitative (interval, ratio)</td>
<td>?</td>
<td>?</td>
</tr>
</tbody>
</table>

Subjective and Objective are often used in research to classify data. Subjective data is qualitative and includes nominal and ordinal levels, while Objective data is quantitative and includes interval and ratio levels.
Subjective – Objective
Quantitative – Qualitative

• Assume you measure 8 times the same attribute of the same entity (A: size [LOC] – B: complexity [?])


Six different Measurement Series
(3 series for attribute ‘A’; 3 series for attribute ‘B’)

### Subjective – Objective

|---|---------------------------------------------|------------------------|

**Guess:**
- Columns 1 to 5 are Quantitative
- Columns 4 & 5 Might be Labels (not Numbers)
Subjective – Objective
Quantitative – Qualitative

• Assume you measure 8 times the same attribute of the same entity (A: size [LOC] – B: complexity [?])


Guess: Columns 2 and 4 are Objective

BUT: What if Column 4 Had value ’high’?
Objective vs. Subjective Measurement

- **Objective Measurement**
  - Usually, the measurement process can be automated
  - (Almost) no random measurement error, i.e., the process is perfectly reliable

- **Subjective Measurement**
  - Human involvement in the measurement process
  - If we repeat the measurement of the same object(s) several times, we might not get exactly the same measured value every time, i.e., the measurement process is not perfectly reliable

**Rule of Thumb:**
Subjective measures have proven to be useful – but if an objective measure is available, then it is (usually) preferable
Procedures for Subjective Measurement

- Subjective Measures usually entail a well-defined Measurement Procedure that precisely describes:
  - How to collect the data (usually via questionnaires on paper or online)
  - How to conduct interviews
  - How to review documents (software artifacts)
  - In which order to assess the dimensions/items of the data collection instrument, etc.

- Examples: ISO9000 Audit, CMMI/SPICE Assessment, Function Points
Objective vs. Subjective Measurement

Examples:

- **Subjective Measurement**
  - Classification of defects into severity classes
  - Function Points (when counted manually)
  - Software Process Assessments

- **Objective Measurement**
  - Lines of Code
  - Cyclomatic Complexity
  - Memory Size
  - Test Coverage

To which category belong …
  - Effort?
  - Time?
  - Defect Count?
Basic Concepts in Subjective Measurement

- **Construct**: A conceptual object that cannot be directly observed and therefore cannot be directly measured (i.e., we estimate the quantity we are interested in rather than directly measure it); for example:
  - User Satisfaction
  - Competence of a Software Engineer
  - Efficiency of a Process
  - Maturity of an Organization

- **Item**: A subjective measurement scale that is used to measure a construct
  - A question on a questionnaire is an item
Dimensionality of Constructs

- Constructs can be one-dimensional or multi-dimensional.
- If a construct is multidimensional, then each dimension covers a different and distinct aspect of the construct.
  - e.g., the different dimensions of customer satisfaction.
Likert Type Scales

• Evaluation-type
  Example:
  “Familiarity with and comprehension of the software development environment”
  ❑ Little
  ❑ Unsatisfactory
  ❑ Satisfactory
  ❑ Excellent

• Frequency-type
  Example:
  “Customers provide information to the project team about the requirements”
  ❑ Never
  ❑ Rarely
  ❑ Occasionally
  ❑ Most of the time

• Agreement-type
  Example:
  “The tasks supported by the software at the customer site change frequently”
  ❑ Strongly Agree
  ❑ Agree
  ❑ Disagree
  ❑ Strongly Disagree
Semantic Differential Scale

Items which include semantic opposites

Example:

“Processing of change requests to existing systems or services: the time that MIS staff takes until responding to change requests received from users of existing computer-based information systems or services.”

Slow □ □ □ □ □ □ □ □ □ Fast
or
Timely □ □ □ □ □ □ □ □ □ Untimely
Assigning numbers to scale responses

• **Likert-Type Scales:**
  - Strongly Agree  -> 1
  - Agree          -> 2
  - Disagree       -> 3
  - Strongly Disagree  -> 4

• **Ordinal Scale**
• **But:**
  Often the distances between the four response categories are approximately (conceptually) equidistant and thus are treated like approximate interval scales.

• **Semantic Differential Scale:**

  Slow □ □ □ □ □ □ □ Fast
  
  1 2 3 4 5 6 7

• **Ordinal scale, but again, often treated as interval scales**
Reliability versus Validity

Assume you measure several times the same attribute of an entity (say, complexity of a code module) and the centre point is the true (but unknown) value.

http://www.uni.edu/chfasoa/reliabilityandvalidity.htm
Reliability versus Validity

Assume you measure several times the same attribute of an entity (say, complexity of a code module) and the centre point is the true (but unknown) value.

Not reliable: too much random bias (noise)
Not valid: too much systematic bias
Reliability Estimation Techniques – Classes

- **Number of administrations** is the number of times that the same object is measured (per observer)
- **Number of instruments** is the number of different but equivalent instruments that would need to be administered

<table>
<thead>
<tr>
<th>Number of Instruments</th>
<th>Number of Administrations (per Observer / Rater)</th>
</tr>
</thead>
<tbody>
<tr>
<td>One</td>
<td>Inter-Rater Internal Consistency</td>
</tr>
<tr>
<td>Two</td>
<td>Test-Retest</td>
</tr>
<tr>
<td></td>
<td>Parallel Forms (immediate)</td>
</tr>
<tr>
<td></td>
<td>Parallel Forms (delayed)</td>
</tr>
</tbody>
</table>

http://www.socialresearchmethods.net/kb/reltypes.php
Inter-Rater Agreement vs. Internal Consistency

• Example

<table>
<thead>
<tr>
<th>Book 1</th>
<th>Book 2</th>
<th>Book 3</th>
<th>Book 4</th>
</tr>
</thead>
</table>

4 Books

<table>
<thead>
<tr>
<th>R1</th>
<th>R2</th>
</tr>
</thead>
</table>

2 Reviewers

<table>
<thead>
<tr>
<th>Quality</th>
<th>1 2 3 4 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Readability</td>
<td>bad</td>
</tr>
<tr>
<td>Suspense</td>
<td>little</td>
</tr>
<tr>
<td>Length</td>
<td>long</td>
</tr>
<tr>
<td>Weight</td>
<td>heavy</td>
</tr>
</tbody>
</table>

1 Instrument – 4 Items
### Inter-Rater Agreement vs. Internal Consistency

#### Example – Data

<table>
<thead>
<tr>
<th>R1:</th>
<th>Book 1:</th>
<th></th>
<th>Book 2:</th>
<th></th>
<th>Book 3:</th>
<th></th>
<th>Book 4:</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Q:</td>
<td></td>
<td>Q:</td>
<td></td>
<td>Q:</td>
<td></td>
<td>Q:</td>
<td></td>
</tr>
<tr>
<td></td>
<td>- R: 2</td>
<td></td>
<td>- R: 4</td>
<td></td>
<td>- R: 2</td>
<td></td>
<td>- R: 4</td>
<td></td>
</tr>
<tr>
<td></td>
<td>- S: 3</td>
<td></td>
<td>- S: 3</td>
<td></td>
<td>- S: 3</td>
<td></td>
<td>- S: 5</td>
<td></td>
</tr>
<tr>
<td></td>
<td>- L: 3</td>
<td></td>
<td>- L: 2</td>
<td></td>
<td>- L: 1</td>
<td></td>
<td>- L: 4</td>
<td></td>
</tr>
<tr>
<td></td>
<td>- W: 3</td>
<td></td>
<td>- W: 2</td>
<td></td>
<td>- W: 2</td>
<td></td>
<td>- W: 2</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>R2:</th>
<th>Book 1:</th>
<th></th>
<th>Book 2:</th>
<th></th>
<th>Book 3:</th>
<th></th>
<th>Book 4:</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Q:</td>
<td></td>
<td>Q:</td>
<td></td>
<td>Q:</td>
<td></td>
<td>Q:</td>
<td></td>
</tr>
<tr>
<td></td>
<td>- R: 3</td>
<td></td>
<td>- R: 3</td>
<td></td>
<td>- R: 2</td>
<td></td>
<td>- R: 4</td>
<td></td>
</tr>
<tr>
<td></td>
<td>- S: 3</td>
<td></td>
<td>- S: 4</td>
<td></td>
<td>- S: 1</td>
<td></td>
<td>- S: 4</td>
<td></td>
</tr>
<tr>
<td></td>
<td>- L: 3</td>
<td></td>
<td>- L: 3</td>
<td></td>
<td>- L: 2</td>
<td></td>
<td>- L: 3</td>
<td></td>
</tr>
<tr>
<td></td>
<td>- W: 3</td>
<td></td>
<td>- W: 2</td>
<td></td>
<td>- W: 2</td>
<td></td>
<td>- W: 2</td>
<td></td>
</tr>
</tbody>
</table>

#### Average Inter-Item Correlation

<table>
<thead>
<tr>
<th>R</th>
<th>S: 0.41</th>
<th>L: -0.11</th>
<th>W: -0.41</th>
</tr>
</thead>
<tbody>
<tr>
<td>Avg</td>
<td>0.15</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

#### Spearman Rank Correlation

\[
(R) = 1 - \frac{6 \sum d^2}{n^3 - n}
\]

<table>
<thead>
<tr>
<th>R</th>
<th>S</th>
<th>Rank[R]</th>
<th>Rank[S]</th>
<th>diff</th>
<th>d^2</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>3</td>
<td>2</td>
<td>4</td>
<td>-2</td>
<td>4</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>7</td>
<td>4</td>
<td>3</td>
<td>9</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>7</td>
<td>7</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>4.5</td>
<td>4</td>
<td>0.5</td>
<td>0.3</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>4.5</td>
<td>7</td>
<td>-2.5</td>
<td>6.3</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>7</td>
<td>7</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

\[
R = \frac{3(21)}{6}\sum d^2
\]
Inter-Rater Agreement vs. Internal Consistency

Example – Data

**Average Inter-Item Correlation**

<table>
<thead>
<tr>
<th></th>
<th>R:</th>
<th>S:</th>
<th>L:</th>
<th>W:</th>
</tr>
</thead>
<tbody>
<tr>
<td>R1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R2</td>
<td>-0.41</td>
<td>0.41</td>
<td>-0.11</td>
<td>0.53</td>
</tr>
</tbody>
</table>

Avg = 0.15

**Inter-rater Agreement (Readability):**

R1: R: 2 – 4 – 2 – 4
R2: R: 3 – 3 – 2 – 4

Fleiss’ Kappa = 0.33
(fair agreement)
Relevant Literature (Selection)

SLR:


Relevant Literature (Selection)

MS:

QS:
• … [parts 2-5]