Systematic secondary studies
TIEJ601 Postgraduate Seminar in Mathematical Information Technology

Antti-Juhani Kaijanaho

Department of Mathematical Information Technology
University of Jyväskylä

September 30, 2014
Definition
A secondary study is a study about some topic that uses published scientific literature (called primary studies) about the same topic as its source of data.

Note
The term tertiary study is sometimes used of secondary studies using other secondary studies as their source of data.
Systematic secondary studies

- deliberately designed
- best current practice followed
- audit trail preserved
- meticulously documented
- threats to validity assessed
- time and effort intensive
Main types

- **Systematic Literature Review (SLR)**
  - very focused research questions
  - of practical relevance
  - synthesizes a result from the primary studies
  - very common in medicine¹
  - subtype: meta-analysis
    - a set of statistical methods for pooling quantitative primary studies’ data
    - often used to designate SLRs that use meta-analysis methods

- **Systematic Mapping Study (SMS)**
  - broad research questions
  - of relevance primarily to research
  - generates a “map” of the literature
  - rarely attempts a synthesis

- (not exhaustive)

¹See e.g. http://summaries.cochrane.org/
Recommended methodological sources

Software Engineering


- See also the CSE recommendations on the next slide

²https://community.dur.ac.uk/ebse/resources/guidelines/Systematic-reviews-5-8.pdf
Recommended methodological sources

Computer Science Education


- see also the SE recommendations on the previous slide


Conducting a systematic secondary study

1. Determine whether a new systematic secondary study is needed.
2. Design the study.
3. Review (and optionally publish) the design.
4. Conduct literature searches.
5. Select primary studies from the literature found using predefined criteria.
6. Assess the quality of the selected studies (optional for mapping studies).
7. Collect and synthesize data.
8. Assess the limitations of the study.
9. Write and publish one or more reports.
Crosscutting concerns

Steps 4–7 should be
▶ designed
▶ piloted beforehand,
▶ executed carefully
▶ documented while in progress and
▶ assessed for reliability afterward.

- Chapter 3 contains an extensive summary of methodological literature (particularly for software engineering).
- Designed in November 2010.
- Completed in May 2014.
- Approved and published in August 2014.

---

Overall design

1. **Scientific Literature**
   - Systematic documented search

2. **Pool of Potentially Relevant Studies**
   - Systematic documented selection
   - Identification of relevant quotes

3. **Selected Studies**

4. **Thematic Model**
   - Thematic Synthesis
   - Coding
   - Codes
   - "..."
Research questions

What scientific evidence is there about the efficacy of particular decisions in programming language design?

1. How much has the efficacy of particular programming language design decisions been empirically studied?

2. Which programming language design decisions have been studied empirically for efficacy?

3. Which facets of efficacy regarding programming language design decisions have been studied empirically?

4. Which empirical research methods have been used in studying the efficacy of particular programming language design decisions?

5. How common are follow-up or replication studies, either by the original researchers or by others?
Searches

Journals

<table>
<thead>
<tr>
<th>Journal</th>
<th>Vols.</th>
<th>Years</th>
<th>Date of search</th>
<th>Yield</th>
</tr>
</thead>
</table>

Proceedings

<table>
<thead>
<tr>
<th>Proc. of</th>
<th>Years</th>
<th>Date of search</th>
<th>Yield</th>
</tr>
</thead>
</table>

Keyword search

Keyword search in Google Scholar, IEEE Xplore, ISI Web of Science, and ScienceDirect performed in 2011 and 2013 yielded 420 candidate publications.

Snowball search

Searching in the references and in the set of citing articles of already selected articles was performed in Spring 2013 and yielded 293 candidate publications.
### Assessing manual search

<table>
<thead>
<tr>
<th></th>
<th>QGS</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>total</td>
<td>contrib.</td>
<td>q.-s.</td>
</tr>
<tr>
<td>ESE</td>
<td>3</td>
<td>1</td>
<td>33 %</td>
</tr>
<tr>
<td>CACM</td>
<td>4</td>
<td>4</td>
<td>100 %</td>
</tr>
<tr>
<td>TOPLAS</td>
<td>9</td>
<td>9</td>
<td>100 %</td>
</tr>
<tr>
<td>LOPLAS</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>IJMMS</td>
<td>14</td>
<td>12</td>
<td>86 %</td>
</tr>
<tr>
<td>IJHCS</td>
<td>2</td>
<td>2</td>
<td>100 %</td>
</tr>
<tr>
<td>PPIG</td>
<td>2</td>
<td>1</td>
<td>50 %</td>
</tr>
<tr>
<td>ISESE</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>ESEM</td>
<td>2</td>
<td>2</td>
<td>100 %</td>
</tr>
<tr>
<td>OOPSLA</td>
<td>14</td>
<td>9</td>
<td>64 %</td>
</tr>
<tr>
<td>SPLASH</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>ECOOP</td>
<td>13</td>
<td>13</td>
<td>100 %</td>
</tr>
<tr>
<td>POPL</td>
<td>3</td>
<td>3</td>
<td>100 %</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>66</td>
<td>56</td>
<td>85 %</td>
</tr>
</tbody>
</table>

QGS = quasi-gold standard: (in this case) the set of relevant publications published in the forum found by any search

contr. = contribution (in the QGS): (in this case) the set of relevant publications published in the forum found by manual search in the forum

q.-s. = quasi-sensitivity: (in this case) the ratio of relevant publications found by manual search in the forum to the number of relevant publications found by any search in the forum
## Assessing automatic searches

<table>
<thead>
<tr>
<th></th>
<th>yield</th>
<th>oa.</th>
<th>QGS</th>
<th>q.-s.</th>
<th>sp.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Google Scholar</td>
<td>8 455</td>
<td>67</td>
<td>16</td>
<td>24 %</td>
<td>1 %</td>
</tr>
<tr>
<td>IEEE Xplore</td>
<td>995</td>
<td>18</td>
<td>2</td>
<td>3 %</td>
<td>18 %</td>
</tr>
<tr>
<td>ScienceDirect</td>
<td>1 022</td>
<td>8</td>
<td>7</td>
<td>11 %</td>
<td>1 %</td>
</tr>
<tr>
<td>Web of Science</td>
<td>20</td>
<td>3</td>
<td>0</td>
<td>0 %</td>
<td>15 %</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>10 492</td>
<td>69</td>
<td>18</td>
<td>27 %</td>
<td>1 %</td>
</tr>
</tbody>
</table>

oa. contrib. = overall contribution: the number of relevant publications found by this search engine in any forum

QGS contrib. = contribution in the quasi-gold standard: the number of relevant publications found by this search engine within the manually searched forums

q.-s. = quasi-sensitivity: the ratio of the size of the QGS contribution to the size of the full QGS (the set of relevant publications found by any search in any manually-searched forum)

sp. = specificity: ratio of overall contribution to the yield
Selection criteria

1. Is this a primary study that attempts to determine the efficacy of a programming language design decision? (If not, skip question 5.)

2. Is this a literature review that attempts to summarize or consolidate research on the efficacy of a programming language design decision? (If not, skip questions 6 and 7.)

3. Can you find a complete written and published report about this study?

4. Is the study reported in English, Finnish or Swedish?

5. Does this primary study present scientific empirical evidence about their claims?

6. Does this secondary study include any primary studies that present scientific empirical evidence?

7. Does this secondary study discuss scientific empirical evidence in the primary studies under review?

EXCLUDE if Q1 and Q2 are both NO or any of Q3–Q7 is NO.
Selection process

Initial search

1515

Search update

248

Exclusion decisions based on on-line metadata

1045

151

Exclusion decisions based on full text

92

26

186 articles provisionally included in the study

Snowballing

293

223

26

68

References and citing articles
Selection validation

Pairwise Cohen $\kappa$

<table>
<thead>
<tr>
<th></th>
<th>AJK-2</th>
<th>VT</th>
<th>VL</th>
<th>TK</th>
</tr>
</thead>
<tbody>
<tr>
<td>AJK-2</td>
<td>0.82 (+0.65 to 0.99)</td>
<td>1.00 (+1.00 to 1.00)</td>
<td>0.62 (-0.10 to 1.00)</td>
<td>0.29 (-0.26 to 0.83)</td>
</tr>
<tr>
<td>VT</td>
<td>0.78 (+0.36 to 1.00)</td>
<td>1.00 (+1.00 to 1.00)</td>
<td>0.62 (-0.10 to 1.00)</td>
<td>0.38 (-0.15 to 0.92)</td>
</tr>
<tr>
<td>VL</td>
<td>0.62 (-0.10 to 1.00)</td>
<td>0.62 (-0.10 to 1.00)</td>
<td>0.62 (-0.10 to 1.00)</td>
<td>0.22 (-0.45 to 0.90)</td>
</tr>
<tr>
<td>TK</td>
<td>0.29 (-0.26 to 0.83)</td>
<td>0.38 (-0.15 to 0.92)</td>
<td>0.22 (-0.45 to 0.90)</td>
<td>0.38 (-0.40 to 1.00)</td>
</tr>
<tr>
<td>$\kappa$</td>
<td>AJK-1</td>
<td>AJK-2</td>
<td>VT</td>
<td>VL</td>
</tr>
</tbody>
</table>

AJK-1 is my original set of decisions ($n = 2056$), AJK-2 is my set of re-examinations ($n = 100$), and VT ($n = 28$), TK ($n = 20$), and VL ($n = 10$) are my three supervisors; the pairwise comparisons use the smaller $n$ of the pair, except between TK and VT ($n = 19$).

Multi-way Fleiss $\kappa$

For all ratings, $\kappa = 0.42$ (95 % CI $-0.19$ to $1.00$, $n = 10$).
For all except TK, $\kappa = 0.77$ (95 % CI $0.02$ to $1.00$, $n = 10$).

Note

Some sources recommend avoiding the Cohen $\kappa$.
Krippendorff $\alpha$ is often recommended as the best choice (not done in my licentiate thesis).\(^6\)

\(^6\)See e. g. Hayes & Krippendorff: *Answering the Call for a Standard Reliability Measure for Coding Data*. Communication Methods and Measures, 1 (1) 77–89. doi:10.1080/19312450709336664.
Synthesis and mapping


1. I read all included studies at least once, in order to “get immersed with the data” (p. 276).

2. I extracted certain categories of information from each paper, as direct quotes (with page references)

3. I abstracted and created a code book for those categories of information as well as any emergent concepts and categories, while simultaneously applying it to the quotes.

4. I explored the resulting codings to determine interesting themes and patterns.

For results, see the licentiate thesis and

Threats to validity

- The searches may not have found all relevant publications.
- The selection process may not have reliably determined which publications are relevant and which are not.
- The coding process may not have reliably encoded the relevant information in the primary studies.
- The thematic synthesis process may have created misleading themes and patterns.
Dissemination plan

Recommended at least (done in this case)

- a peer-reviewed detailed report, either a thesis or as a technical report (thesis finished)
- academic article (journal article in preparation)
- trade article if relevant (no plans at this time)
- reports for other stakeholders if relevant (not relevant)
Personal observations

- properly done a multi-person-year undertaking
  - good idea to have a team of 3–6 researchers
- most abstracts are almost useless
  - I recommend adopting structured abstracts

---

7See e.g. https://community.dur.ac.uk/ebse/abstracts.php
DISCUSSION TIME