Supplementary Materials for the paper "*Meta-analysis for Families of Experiments in Software Engineering: A Systematic Review and Reproducibility and Validity Assessment*"

Barbara Kitchenham, Lech Madeyski, and Pearl Brereton

May 1, 2019

1 Systematic Review Process Details

1.1 Selection procedures

The first author (BAK) applied the inclusion and exclusion criteria to the identified candidate primary studies. The third author (PB) checked the application of the inclusion/exclusion criteria to each candidate primary study. The single disagreement during the search and selection process was resolved by discussion.

2 SR primary study search process

Our main search process was an automated search using SCOPUS because it indexes all five journals.

We initially validated our automated search strings by checking that they found three studies conforming with our inclusion criteria that we were aware of before we began this SR:

- (Scanniello et al. 2014).

- (Abrahao et al. 2013).
- (Laitenberger et al. 2001).

Further validation of the search process involved checking the SCOPUS searches against similar automated searches performed on the DBLP database. This was intended to refine our search string(s). Finally, the search strings applied to SCOPUS were validated by applying equivalent searches to the Semantic Scholar database. The results of this process are reported in subsequent sections.

2.1 Search strings

All the searches and search validation discussed in this and the next section took place between November 7th and November 19th 2017. All the journals except TSE published all the volumes of their journals for 2017 prior to November 7th, therefore, we searched the final volumes of TSE manually to check for any additional relevant papers.

Our initial search string was:

> *SR1: TITLE-ABS-KEY ("family of experiment*") AND SUBJAREA (comp) AND (LIMIT-TO (DOCTYPE , "ar")) AND LIMIT-TO(EXACTSRCTITLE, "Empirical Software Engineering") AND LIMIT-TO(EXACTSRCTITLE, "Journal Of Systems And Software") AND LIMIT-TO(EXACTSRCTITLE, "IEEE Transactions On Software Engineering") AND LIMIT-TO(EXACTSRCTITLE, "Information And Software Technology") AND LIMIT-TO(EXACTSRCTITLE, "ACM Transactions On Software Engineering And Methodology")*

This search identified 20 candidate primary studies, of which BAK assessed 10 to be suitable for inclusion in the SR. The search found two of the papers we were already aware of, i.e., (Scanniello et al. 2014) and (Abrahao et al. 2013), but missed (Laitenberger et al. 2001) which did not use the term "*family of experiments*".

To check the SCOPUS results, the second author (LM) performed the following title only search on the DBLP database: `http://dblp.uni-trier.de/search?q=family%20of%20experiments`

This search found 25 candidate primary studies. It included nine papers not found by the SCOPUS search. Eight papers were published

in sources other than the five designated journals. In addition, the search found one candidate paper that ought to have been found by the SCOPUS search (i.e., Muñoz et al. 2010). Muñoz et al. was published in IST and included the term "family of experiments", so it should have been found. However, it was not found because SCOPUS has set the document type of this paper to "Conference paper". Therefore we refined the string, removing unnecessary restrictions, as shown here:

> *SR1: TITLE-ABS-KEY ("family of experiment*") AND (LIMIT-TO (EXACTSRCTITLE,"Empirical Software Engineering ") OR LIMIT-TO (EXACTSRCTITLE,"Journal Of Systems And Software") OR LIMIT-TO (EXACTSRCTITLE,"IEEE Transactions On Software Engineering") OR LIMIT-TO (EXACTSRCTITLE,"Information And Software Technology") OR LIMIT-TO (EXACTSRCTITLE,"ACM Transactions On Software Engineering And Methodology"))*

This search identified 22 candidate primary studies[1], including Muñoz et al. 2010. It also found another candidate primary study (Gonzalez-Huerta et al. 2015), which was also categorised as a conference paper by SCOPUS. At this point BAK had identified 11 potential primary studies. PB checked all the papers and agreed with all exclusions and inclusions.

In order to address the problem of primary studies that did not explicitly call themselves families of experiments, e.g. (Laitenberger et al. 2001), we conducted another SCOPUS search using the following search string:

> *SR2: TITLE-ABS-KEY ("replicat*") AND TITLE-ABS-KEY ("meta-analysis") AND (LIMIT-TO(EXACTSRCTITLE, "Empirical Software Engineering") OR LIMIT-TO(EXACTSRCTITLE, "Journal Of Systems And Software") OR LIMIT-TO(EXACTSRCTITLE, "IEEE Transactions On Software Engineering") OR LIMIT-TO(EXACTSRCTITLE, "Information And Software Technology") OR LIMIT-TO(EXACTSRCTITLE, "ACM Transactions On Software Engineering And Methodology"))*

[1]The results are the same if the LIMIT-TO(EXACTSRTITLR, "Journal name") is replaced by ISSN(ISSNnumber), for example ISSN(1049331X) for ESE.

We limited the search to papers that included the term “meta-analysis" in the title, abstract or keywords to avoid excessive numbers of false positives.

The search found 7 papers including 4 that were found in the first search and Laitenberger et al. 2001. BAK judged the remaining two papers, i.e., (Pfahl et al. 2004) and (Acuña et al. 2015), to be potentially relevant to our SR, since they both used meta-analysis to determine the aggregate standardised effect sizes from three experiments. After checking the full text of the two new candidate primary studies, PB agreed with the decision to include them.

Hence, at the completion of the main search and selection process we identified a total of 14 primary studies for inclusion in our systematic review.

2.2 Search Validation

Our SCOPUS searches were validated by performing semantically equivalent searches on the Semantic Scholar digital library (`https://www.semanticscholar.org`).

The combination of the two semantic scholar searches found all but one of the 14 primary studies found by the SCOPUS searches and did not find any new primary studies. These results suggest that we have found most (if not all) of the papers that are relevant to our systematic review, and were published in our five designated journals.

After we had completed our search process, we found that Santos et al. 2018 had undertaken a mapping study of families experiments. Their study reported 15 families of experiments that meta-analyzed effect sizes. Two of those papers were not published in the five sources we searched, so were not considered relevant for our systematic review. We found all the other papers that they found, plus one additional paper, i.e., (Morales et al. 2016).

2.3 Completion of the Search and Selection Processes

Manual checking of TSE for November and December 2017 performed by LM and checked by BAK found no new primary studies. Thus, at the end of the process we identified 14 primary studies for inclusion in the SR.

3 Data extraction

This section describes both the data we extracted from each study and how the data extraction was conducted.

3.1 Extracted data

The data we extracted from our primary studies was broken up into the following categories:

1. Primary study identification information. This included a study identifier and the study citation.

2. Information relating to the goals of the study (both the study as a whole and those of the individual experiments). This included the stated goals, the software engineering methods being investigated and the hypotheses being tested.

3. Information needed to understand the study, the individual experiments and differences between the individual experiments. This included the number and type of participants in each experiment, the number and type of tasks they were asked to perform, and the software engineering materials they used to perform those tasks.

4. Data related to good experimental practice. This included whether a pilot study was performed, whether power analyses was reported, whether the software materials were available on line, whether the raw data was available on line.

5. Data related to our research goals:

    - To address RQ1 and RQ2, this included the statistical design reported for each experiment and any justification for the design, and the meta-analysis process that the authors used to aggregate effect sizes, in particular, the description of the process reported by the authors.
    - To address RQ3 and RQ4, we extracted the descriptive statistics for each experiment in the family from which we could attempt to reproduce the meta-analysis process used by the

authors. We also extracted the effect sizes reported for the individual studies together with the aggregated measures with any confidence intervals or probability values reported. We could therefore compare the meta-analysis results we obtained with those reported in the primary studies.

3.2 Data extraction

BAK extracted the data from each study. Although RQ1 and RQ2 were aimed primarily at identifying effect size and meta-analysis process, we also identified the experimental design from the description of the experimental design reported in the primary study.

3.3 Data extraction validation

PB independently checked the extracted data and identified 49 disagreements in 1970 data items (i.e., 2.49%). In all cases PB's revised values were accepted as correct.[2]

BAK copied the data needed for reproducibility analyses (i.e., the mean and standard deviations of the data reported as descriptive statistics) to an R script. LM checked 238 values and found 7 errors. BAK checked a further 94 items and found 2 errors. All 9 errors were corrected. In addition, BAK incorrectly extracted the number of observations rather than the number of participants for 5 papers which required 30 subsequent corrections.

3.4 Final Selection of Primary Studies

An issue that arose during data extraction was that although the paper by Acuña et al. 2015 reported results from three studies using a correlation effect size and was referred to by the authors as a quasi-experiment, it seemed, in fact, to be an observational study. In particular, it did not compare any software engineering methods. After discussion, we agreed

[2]BAK made all the necessary corrections and PB checked them. She found three corrections that had been missed. These were all the same correction that followed on from a preceding correction.

to omit this study from our SR. Thus, addressing RQ1, we included only 13 primary studies in our data analysis process.

4 The Variation among Experiments in each Family

Table 1 reports the extent to which the individual studies vary one from another.

Table 1: Experimental Design Variation Among Families

ID	Num Exp Designs	Num SW Documents	Num Participant Types	Num Institutes
Study 11	1	1	2	2
Study 5	1	1	2	3
Study 8	1	1	2	2
Study 9	1	1	1	2
Study 2	1	1	2	2
Study 1	1	2	4	3
Study 4	1	1	2	1
Study 6	1	1	1	1
Study 7	1	1	2	3
Study 10	1	1	1	2
Study 3	2	3	3	5
Study 14	1	1	2	3
Study 13	1	3	1	1

All counts start from 1, so a value of 1 for a factor for a primary study means there were no changes to that factor. The number of software documents[3] is counted with respect to pairs of documents, so a change

[3]I.e., the software materials or artefacts or programs used to undertake each experimental task. Questionnaires or forms completed by the participants to provide the outcome measures are not included in these counts.

can apply to one or both of the documents that participants used to perform the experimental tasks. However, changes to the documents in order to implement them in two different forms (to represent two different procedures or methods) were not counted as changes to the document pairs.

References

Abrahao, Silvia, Carmine Gravino, Emilio Insfran Pelozo, Giuseppe Scanniello, and Genoveffa Tortora (2013). "Assessing the Effectiveness of Sequence Diagrams in the Comprehension of Functional Requirements: Results from a Family of Five Experiments". In: *IEEE Transactions on Software Engineering* 39.3, pp. 327–342.

Acuña, Silvia T., Marta N. Gómez, Jo E. Hannay, Natalia Juristo, and Dietmar Pfahl (2015). "Are team personality and climate related to satisfaction and software quality? Aggregating results from a twice replicated experiment". In: *Information and Software Technology* 57.1, pp. 141–156.

Gonzalez-Huerta, Javier, Emilio Insfrán, Silvia Mara Abrahão, and Giuseppe Scanniello (2015). "Validating a model-driven software architecture evaluation and improvement method: A family of experiments". In: *Information and Software Technology* 57, pp. 405–429.

Laitenberger, Oliver, Khaled El Emam, and Thomas G. Harbich (2001). "An internally replicated quasi-experimental comparison of checklist and perspective-based reading of code documents". In: *IEEE Transactions on Software Engineering* 27.5, pp. 387–418.

Morales, José Miguel, Elena Navarro, Pedro Sánchez-Palma, and Diego Alonso (2016). "A family of experiments to evaluate the understandability of TRiStar and i* for modeling teleo-reactive systems". In: *Journal of Systems and Software* 114, pp. 82–100.

Muñoz, Lilia, Jose-Norberto Mazón, and Juan Trujillo (2010). "A family of experiments to validate measures for UML activity diagrams of ETL processes in data warehouses". In: *Information and Software Technology* 52.11, pp. 1188–1203.

Pfahl, Dietmar, Oliver Laitenberger, Günther Ruhe, Jörg Dorsch, and Tatyana Krivobokova (2004). "Evaluating the learning effectiveness

of using simulations in software project management education: results from a twice replicated experiment". In: *Information and Software Technology* 46.2, pp. 127–147.

Santos, Adrian, Omar S. Gómez, and Natalia Juristo (2018). "Analyzing Families of Experiments in SE: A Systematic Mapping Study". In: *CoRR* abs/1805.09009. arXiv: 1805.09009. URL: http://arxiv.org/abs/1805.09009.

Scanniello, Giuseppe, Carmine Gravino, Marcela Genero, Jose' A. Cruz-Lemus, and Genoveffa Tortora (Apr. 2014). "On the Impact of UML Analysis Models on Source-code Comprehensibility and Modifiability". In: *ACM Transactions on Software Engineering and Methodology* 23.2, 13:1–13:26. DOI: 10.1145/2491912.