Replications Studies in Software Engineering Research

Tamer Abdou, Ph.D.

Industrial Engineering

Ryerson University

Toronto, Ontario

Canada

Wednesday, July 11, 2018
1 Introduction & Terminologies

2 A Set of Guidelines to Follow
   - Information about the original study
   - Information about the replication
   - Comparison of results to original
   - Conclusion Across Studies

3 Types of Replication Studies
   - Same experiment & Same objects
   - Different experiment & Same objects
   - Same experiment & Different objects
Replication Study

- A study that involves sharing information/knowledge so as to ensure consistency between redundant resources, such as software or hardware components¹.

- A study based on the design, methodology and results of previously published research papers².

Replication Study

- A study that involves **sharing information/knowledge** so as to ensure **consistency** between **redundant resources**, such as software or hardware components\(^1\).
- A study based on the **design, methodology** and **results** of **previously** published research papers\(^2\).

Do We Need to Replicate Studies in SE?

- Concerns about the **reliability** of empirical research results are fast becoming endemic and software engineering is no exception.
- **False discoveries** and how likely published experiments report erroneous results.
- Researchers questioned the **prevalence** of reported p-values.
- Concerns about the **variability of results** depending upon which research team performs the work.
- Some studies are **selectively published** based on preferences for particular results.
- There is both a **low probability** of discovering a **true effect** and the parameter of interest has high variance.

Do We Need to Replicate Studies in SE?

- Concerns about the **reliability** of empirical research results are fast becoming endemic and software engineering is no exception.

- **False discoveries** and how likely published experiments report erroneous results.

- Researchers questioned the **prevalence** of reported p-values

- Concerns about the **variability of results** depending upon which research team performs the work.

- Some studies are **selectively published** based on preferences for particular results.

- There is both a **low probability** of discovering a true effect and the parameter of interest has high variance.

Do We Need to Replicate Studies in SE?

- Concerns about the **reliability** of empirical research results are fast becoming endemic and software engineering is no exception.

- **False discoveries** and how likely published experiments report erroneous results.

- Researchers questioned the **prevalence** of reported **p-values**

- Concerns about the **variability of results** depending upon which research team performs the work.

- Some studies are **selectively published** based on preferences for particular results.

- There is both a **low probability** of discovering a **true effect** and the parameter of interest has high variance.

Do We Need to Replicate Studies in SE?

- Concerns about the **reliability** of empirical research results are fast becoming endemic and software engineering is no exception.
- **False discoveries** and how likely published experiments report erroneous results.
- Researchers questioned the **prevalence** of reported **p-values**
- Concerns about the **variability of results** depending upon which research team performs the work.
- Some studies are **selectively published** based on preferences for particular results.
- There is both a **low probability** of discovering a **true effect** and the parameter of interest has high variance.

Do We Need to Replicate Studies in SE?

- Concerns about the **reliability** of empirical research results are fast becoming endemic and software engineering is no exception.
- **False discoveries** and how likely published experiments report erroneous results.
- Researchers questioned the **prevalence** of reported **p-values**
- Concerns about the **variability of results** depending upon which research **team performs the work**.
- Some studies are **selectively published** based on preferences for particular results.
- There is both a **low probability** of discovering a **true effect** and the parameter of interest has high variance.

Do We Need to Replicate Studies in SE?

- Concerns about the **reliability** of empirical research results are fast becoming endemic and software engineering is no exception.
- **False discoveries** and how likely published experiments report erroneous results.
- Researchers questioned the **prevalence** of reported **p-values**
- Concerns about the **variability of results** depending upon which research team performs the work.
- Some studies are **selectively published** based on preferences for particular results.
- There is both a **low probability** of discovering a **true effect** and the parameter of interest has high variance.

Evolution of Replications over Years

1994-2003  An average of 4.1 studies published per year.
2004-2009  An average of 11.7 studies published per year.
2004-2012  An average of 24.3 studies published per year.

Topics of Interest

**PROMISE’18** Replication and repeatability of previous work using predictive modelling and data analytics in software engineering

*International conference on Predictive Models and Data Analytics in Software Engineering*

**JSS’18** Replication of empirical studies and families of studies.

*Journal of Systems and Software*

**ESEM’18** Replication of software engineering studies.

*Empirical Software Engineering and Measurement*
Guidelines to follow

1. Information about the original study
2. Information about the replication
3. Comparison of results to original
4. Conclusion Across Studies

1- Research Questions

A description of the research question(s) that was the basis for the original design.
2- Participants

The number of participants and any relevant characteristics of the participants.
3- Design

A graphical (or textual) description of the experimental design.
4- Artifacts

A description of and/or links to the artifacts used.
5- **Context variables**

Any important context variables that affected the design of the study or interpretation of the results
6- Summary of results

A brief overview of the major findings
1- Motivation

a description of why the replication was conducted.
2- Level of interaction

The level of interaction the replicators had with the original experimenter should be reported.
3- Changes to the original experiment

Any changes made to the design, participants, artifacts, procedures, data collected and/or analysis techniques should be discussed here.
1- Similarities in results

Replication results that supported results from the original study.
2- Differences in results

Results from the replication that did not coincide with the results from the original study.
The authors should provide a discussion of the current state of knowledge.
1- Same experiment & Same objects

**Goal:** Evaluating the certainty of current knowledge (i.e., confirming or disputing previous results).
1- **Same experiment & Same objects**

Example (Original Study - Deeper Model)

![Diagram showing Model Building Phase and Prediction Phase](image)

1- Same experiment & Same objects

Example (Original Study - TLEL Model)

1- Same experiment & Same objects

Example (Same Datasets)

<table>
<thead>
<tr>
<th></th>
<th>Period</th>
<th>The total number of changes</th>
<th>Average LOC</th>
<th># of modified files</th>
<th># of changes per day</th>
<th># dev. per file</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>File Change</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bugzilla</td>
<td>08/1998 - 12/2006</td>
<td>4,620</td>
<td>389.8</td>
<td>37.5</td>
<td>2.3</td>
<td>1.5</td>
</tr>
<tr>
<td>Columbia</td>
<td>11/2002 - 07/2006</td>
<td>4,455</td>
<td>125.0</td>
<td>149.4</td>
<td>6.2</td>
<td>3.3</td>
</tr>
<tr>
<td>Eclipse JDT</td>
<td>05/2001 - 12/2007</td>
<td>35,386</td>
<td>260.1</td>
<td>71.4</td>
<td>4.3</td>
<td>14.7</td>
</tr>
<tr>
<td>Eclipse Platform</td>
<td>05/2001 - 12/2007</td>
<td>64,250</td>
<td>231.6</td>
<td>72.2</td>
<td>4.3</td>
<td>26.7</td>
</tr>
<tr>
<td>Mozilla</td>
<td>01/2000 - 12/2006</td>
<td>98,275</td>
<td>360.2</td>
<td>106.5</td>
<td>5.3</td>
<td>38.9</td>
</tr>
<tr>
<td>PostgreSQL</td>
<td>07/1996 - 05/2010</td>
<td>20,431</td>
<td>563.0</td>
<td>101.3</td>
<td>4.5</td>
<td>4.0</td>
</tr>
<tr>
<td>OSS-Median</td>
<td></td>
<td>27,909</td>
<td>310.1</td>
<td>86.7</td>
<td>4.4</td>
<td>9.4</td>
</tr>
<tr>
<td>C-1</td>
<td>10/2000 - 12/2009</td>
<td>4,096</td>
<td>-</td>
<td>16.4</td>
<td>2.0</td>
<td>1.2</td>
</tr>
<tr>
<td>C-2</td>
<td>10/2000 - 12/2009</td>
<td>9,277</td>
<td>-</td>
<td>19.2</td>
<td>2.4</td>
<td>2.8</td>
</tr>
<tr>
<td>C-3</td>
<td>07/2002 - 12/2009</td>
<td>3,586</td>
<td>-</td>
<td>16.6</td>
<td>2.0</td>
<td>1.3</td>
</tr>
<tr>
<td>C-4</td>
<td>12/2003 - 12/2009</td>
<td>5,182</td>
<td>-</td>
<td>12.9</td>
<td>1.8</td>
<td>2.4</td>
</tr>
<tr>
<td>C-5</td>
<td>10/1982 - 12/1995</td>
<td>10,961</td>
<td>303.0</td>
<td>39.0</td>
<td>4.8</td>
<td>2.3</td>
</tr>
<tr>
<td>COM-Median</td>
<td></td>
<td>5,182</td>
<td>-</td>
<td>16.6</td>
<td>2.0</td>
<td>2.3</td>
</tr>
</tbody>
</table>

### Example 1 (Replication Study)

<table>
<thead>
<tr>
<th>Project</th>
<th>Deeper Original</th>
<th>Deeper Replicated</th>
<th>TLEL Original</th>
<th>TLEL Replicated</th>
<th>DSL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bugzilla</td>
<td>0.6292</td>
<td>0.6348</td>
<td>0.6850</td>
<td>0.6722</td>
<td>0.6730</td>
</tr>
<tr>
<td>Columba</td>
<td>0.5606</td>
<td>0.5641</td>
<td>0.6065</td>
<td>0.6050</td>
<td>0.6090</td>
</tr>
<tr>
<td>JDT</td>
<td>0.3779</td>
<td>0.3762</td>
<td>0.4194</td>
<td>0.4125</td>
<td>0.4233</td>
</tr>
<tr>
<td>Mozilla</td>
<td>0.2215</td>
<td>0.2127</td>
<td>0.2625</td>
<td>0.2561</td>
<td>0.2582</td>
</tr>
<tr>
<td>Platform</td>
<td>0.3822</td>
<td>0.3910</td>
<td>0.4471</td>
<td>0.4381</td>
<td>0.4425</td>
</tr>
<tr>
<td>PostgreSQL</td>
<td>0.5509</td>
<td>0.5485</td>
<td>0.6052</td>
<td>0.5958</td>
<td>0.5994</td>
</tr>
<tr>
<td><strong>Average</strong></td>
<td><strong>0.4537</strong></td>
<td><strong>0.4546</strong></td>
<td><strong>0.5043</strong></td>
<td><strong>0.4966</strong></td>
<td><strong>0.5009</strong></td>
</tr>
</tbody>
</table>

2- Different experiment & Same objects

**Goal:** Improving the original model and reduce the internal threats to validity (minimize systematic error)
2- Different experiment & Same objects

Example (Different Model - Same Datasets)

2- Different experiment & Same objects

Example (Replication Study - Same Datasets)

<table>
<thead>
<tr>
<th>Project</th>
<th>Deeper Original</th>
<th>Deeper Replicated</th>
<th>TLEL Original</th>
<th>TLEL Replicated</th>
<th>DSL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bugzilla</td>
<td>0.6292</td>
<td>0.6348</td>
<td>0.6850</td>
<td>0.6722</td>
<td>0.6730</td>
</tr>
<tr>
<td>Columba</td>
<td>0.5606</td>
<td>0.5641</td>
<td>0.6065</td>
<td>0.6050</td>
<td>0.6090</td>
</tr>
<tr>
<td>JDT</td>
<td>0.3779</td>
<td>0.3762</td>
<td>0.4194</td>
<td>0.4125</td>
<td>0.4233</td>
</tr>
<tr>
<td>Mozilla</td>
<td>0.2215</td>
<td>0.2127</td>
<td>0.2625</td>
<td>0.2561</td>
<td>0.2582</td>
</tr>
<tr>
<td>Platform</td>
<td>0.3822</td>
<td>0.3910</td>
<td>0.4471</td>
<td>0.4381</td>
<td>0.4425</td>
</tr>
<tr>
<td>PostgreSQL</td>
<td>0.5509</td>
<td>0.5485</td>
<td>0.6052</td>
<td>0.5958</td>
<td>0.5994</td>
</tr>
<tr>
<td>Average</td>
<td>0.4537</td>
<td>0.4546</td>
<td>0.5043</td>
<td>0.4966</td>
<td>0.5009</td>
</tr>
</tbody>
</table>

3- Same experiment & Different objects

**Goal:** Identifying limitations to the generality of the conclusions (or to problems with the objects).
3- Same experiment & Different objects

Example (Original Study)

3- Same experiment & Different objects

Example (Replication Study - Different Datasets)

Introduction & Terminologies
A Set of Guidelines to Follow
Types of Replication Studies

- Same experiment & Same objects
- Different experiment & Same objects
- Same experiment & Different objects

Thank you!