On the Difficulty of Replicating Human Subject Studies in Software Engineering

Jonathan Lung, Jorge Aranda, Steve Easterbrook and Greg Wilson, ICSE 2008, ACM
On the Difficulty of Replicating Human Subject Studies in Software Engineering
Replication

- Replication is one of the main principles of the scientific method
- Distinction between literal and theoretical replication

<table>
<thead>
<tr>
<th>literal</th>
<th>theoretical</th>
</tr>
</thead>
<tbody>
<tr>
<td>* Come close enough to original experiment to directly compare results.</td>
<td>* Investigate scope of underlying theory</td>
</tr>
<tr>
<td>* Show that same results hold under same conditions</td>
<td>* Show that predictably (dis)similar results hold when conditions are systematically altered.</td>
</tr>
</tbody>
</table>
Replications are Rare

- Lack of information in published reports
  - Lab-packages are possible solution
- Less interesting than novel research
- Perceived to be harder to publish
- Unclear how to assess the cost-benefit trade-off for conducting replications
On the Difficulty of Replicating Human Subject Studies in Software Engineering
Human Subjects

- Human subject studies have highly variable outcomes
  - Good experimental design can eliminate some of the threats to validity (e.g. double-blind trials)
  - Research strategies usually consist of a series of studies. Replication of earlier studies, improved designs or different research method
On the Difficulty of Replicating Human Subject Studies in Software Engineering
Replication in Software Engineering

- SE involves a lot of cognitive and social processes. Leads to inevitable threats to validity
- Creative processes lead to large variations in answers
- Difficult to acquire participants:
  - Skilled personnel may be difficult/expensive to attract
  - Only small subset may be suitable due to variety of tools and languages
- Considered one of the barriers to evidence-based SE (cp. Psychology)
The Camel has Two Humps
The Camel has Two Humps (D&B)

- Unpublished ("stylistic flaws") paper by Saeed Dehnadi and Richard Bornat, Middlesex University, UK
  
  http://www.eis.mdx.ac.uk/research/PhDArea/saeed/paper1.pdf

- What constitutes programming aptitude?
- Previous research was disappointing: grades, mathematics ability, age, sex etc. are poor indicators
- Hypothesis: Usage of mental models allows predicting programming aptitude
Experiment

• 61 Students, no prior programming experience

• Two tests in an introductory programming course:
  • 1\textsuperscript{st} prior to any teaching
  • 2\textsuperscript{nd} after teaching about assignments & sequence (after two weeks)
<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample Question</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>1. Read the following statements and tick the box next to the correct answer in the next column.</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>int a = 10;</td>
<td></td>
<td></td>
</tr>
<tr>
<td>int b = 20;</td>
<td></td>
<td></td>
</tr>
<tr>
<td>a = b;</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>The new values of a and b are:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>□ a = 30   b = 0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>□ a = 30   b = 20</td>
<td></td>
<td></td>
</tr>
<tr>
<td>□ a = 20   b = 0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>□ a = 20   b = 20</td>
<td></td>
<td></td>
</tr>
<tr>
<td>□ a = 10   b = 10</td>
<td></td>
<td></td>
</tr>
<tr>
<td>□ a = 10   b = 20</td>
<td></td>
<td></td>
</tr>
<tr>
<td>□ a = 20   b = 10</td>
<td></td>
<td></td>
</tr>
<tr>
<td>□ a = 0    b = 10</td>
<td></td>
<td></td>
</tr>
<tr>
<td>□ If none, give the correct values:</td>
<td>a =</td>
<td>b =</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Use this column for your rough notes please</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

[D&B]
Mental Models

- How we think about an instruction, e.g. \( a = b; \) can be interpreted in different ways:
  - Value moves from right to left (\( a := b, b := 0 \))
  - Right-hand value extracted and added to left-hand value (\( a := a + b, b := 0 \))
  - Value is copied from right to left (\( a := b, \) “correct”)
Results

- Three groups
  - **Consistent**: 44% of subjects used the same mental model for most (80%) of the questions
  - **Inconsistent**: 39% used different models for different questions
  - **Blank**: 8% refused to answer most of the questions
Results

• Correlation with the exam results (consistent: black, inconsistent/blank: white)
Claim / Speculation

- “We can predict success or failure [in an introductory programming course] even before students have had any contact with any programming language with very high accuracy”
Towards Replication
Why this study?

- Surprising results
- Experiment appears to be sound
- Experiment seems straightforward
- Materials available on website
Experimental Replication

- Authors set out to perform literal replication
- Inevitable changes accumulated. Changes had to be justified
  - Trivial changes: location, recruitment method
  - More serious changes: Instructor was not experimenter, course requirements, test was only administered once, deterministic scoring
Analysis Replication

- Trivial changes: data was compared to original study, blank group was not included in analysis
- Additional statistics
  - Test for self-selection: data is suspect
  - Analyzed correlation between consistency and being above the median (instead of just “passed the exam”)

Analysis of the Replication

- Operationalization of “success” to mean “passed” in D&B is critical
  - Differences between universities make measurement meaningless (Middlesex: 50% fail, Toronto: only 12.9%)
  - Relative measurement more suitable.
    - Proposed: comparing those who do better than the median to those who do worse.
    - Being consistent has no significant correlation with being above or below the median.
    - Also, no difference in the avg. marks of the two groups.

(Operationalization: defining fuzzy concept to make it measurable)
Analysis of the Replication

- Operationalization of “inconsistent” problematic
  - D&B grouped the blank and inconsistent groups. No justification given.

- Threshold for assessing consistency in D&B arbitrary.
  - No significant correlation between degree of consistency and final mark.
Additional Analysis

- Alternative theory: most people from the consistent group are using the Java mental model.
  - Such a group exists, however there is also a group that is consistent with an alternate model.
  - The Java-consistent group does not score better than the inconsistent group, but significantly better than the alternately-consistent group.
- Possible explanation: inconsistent group adapts model, is more flexible when it comes to learning.
Reducing Threats to Validity

- Eliminated the experimenter-expectancy effect
- Deterministic scoring algorithm for responses instead of subjective determination in D&B
- Possibly introduced new threats
  - Students may have downplayed their programming experience (to avoid harder courses)
Observations

- Observations should not be restricted to (dis)confirming results of the replicated experiment, e.g:
  - Some participants may have revised their models facing more complex problems
  - Some participants generated models consciously (using comments). Significance of this is unknown
Summary

- Replication yielded opposite results of original experiment (even with generous interpretation of hypothesis)
- However, the results of D&B are highly unlikely to have occurred by chance. Replication does not imply that results of D&B are wrong!
Summary (Replication)

- No strict comparison was possible. Authors were forced to reiterate upon original lab package
- Literal replication was chosen as an “easy first step” and turned out to be complicated with little results about the underlying theory

but...

- Flaws in the original experiment were identified
- Further research questions were postulated
Important Lessons

- Replicating seemingly straightforward experiments requires acquisition of considerable amount of tacit knowledge
- Seemingly simple instrument may be difficult to apply uniformly
- Attempting to explain differences is fruitful exercise
- Each replication suffers from different set of contextual issues

“Knowledge gain seems modest given the effort we invested”
Review

- Well written, convincing paper
- Paper should probably be divided in two, mix between replication and meta level at times confusing
- Does not follow its own advice
  - Fishing for results
BACKUP
Open Questions

• Very few experience reports, many unanswered questions:
  • Involvement of original research team?
  • Involvement vs. independence
  • Original design vs. Improvements
  • How do variations matter?
  • ...
Mental Models (for $a = b$)

1. Value moves from right to left ($a := b; b := 0$ – third line in figure 1).
2. Value copied from right to left ($a := b$ – fourth line of figure 1, and the ‘correct’ answer).
3. Value moves from left to right ($b := a; a := 0$ – eighth line of figure 1).
4. Value copied from left to right ($b := a$ – fifth line of figure 1, and a reversed version of the ‘correct’ answer).
5. Right-hand value added to left ($a := a+b$ – second line of figure 1).
6. Right-hand value extracted and added to left ($a := a+b; b := 0$ – first line of figure 1).
7. Left-hand value added to right ($b := a+b$ – omitted in error).
8. Left-hand value extracted and added to right ($b := a+b; a:=0$ – omitted in error).
9. Nothing happens (sixth line of figure 1).
10. A test of equality: nothing happens (fourth and fifth lines of figure 1).
11. Variables swap values (seventh line in figure 1).
Inevitable Changes

<table>
<thead>
<tr>
<th>Original</th>
<th>U. of Toronto Replication</th>
<th>Primary reason(s) for difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Participants drawn from courses at Barnet College and Middlesex University [5].</td>
<td>Participants drawn from a course at the University of Toronto.</td>
<td>Not feasible for us to run the experiment in the UK.</td>
</tr>
<tr>
<td>Prior mathematics experience not explicitly required.</td>
<td>All participants had grade 12 mathematics or equivalent.</td>
<td>The course requirements at the University of Toronto include this as a pre-requisite.</td>
</tr>
<tr>
<td>Some participants were contacted, interviewed, and tested prior to taking their first programming course [5].</td>
<td>Participants were not tested until the second week of classes and were not interviewed at all.</td>
<td>The interviews took place at Barnet College due to local protocol concerning admitting students. Final lists of students in courses at the University of Toronto are not available before classes commence.</td>
</tr>
<tr>
<td>The predictive test was administered twice to each participant [5].</td>
<td>The predictive test was administered once to participants.</td>
<td>We could not administer the test before the participants had started programming. Administering the test twice after one week would have likely been minimally informative and discouraging to participants.</td>
</tr>
<tr>
<td>The course from which participants were drawn were taught or tutored by an experimenter [5].</td>
<td>None of the participants were taught or tutored by anyone affiliated with the experiment.</td>
<td>Requirement for informed consent preclude the course instructors from endorsing the study. Also, such involvement could be a possible source of error.</td>
</tr>
<tr>
<td>Recruitment method unknown.</td>
<td>Participants were given a chance to win $50 DVD gift certificates.</td>
<td>A draw for gift certificates was within our means and an often-used method at the University of Toronto.</td>
</tr>
<tr>
<td>Participants were asked what A-Level courses or equivalents were taken [5].</td>
<td>Participants were asked what courses were taken to the highest high school level.</td>
<td>Participants in Toronto would be unfamiliar with the UK GCE system.</td>
</tr>
<tr>
<td>Responses by participants were coded using a subjective system [5,6].</td>
<td>Responses by participants were coded using an automated tool.</td>
<td>Eliminating subjectivity increases reproducability. Since both produced the same result on some test data, results should be comparable. Further, using an automated approach improved reliability, speed, and facilitated additional analysis.</td>
</tr>
</tbody>
</table>
### Changes

<table>
<thead>
<tr>
<th>Original</th>
<th>U. of Toronto Replication</th>
<th>Reason for difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tests for self-selection not performed/shown.</td>
<td>Tests for self-selection performed.</td>
<td>Testing for self-selection is a relatively easy way to test data validity and is good practice.</td>
</tr>
<tr>
<td>Transformations of data not performed/shown.</td>
<td>Recoding responses of HTML-only experience as having no experience. Exclusion of data of participants who claimed prior programming experience. Removal of data of participants who had no final mark available.</td>
<td>HTML is not a programming languages and those with programming experience are not part of the population of interest. Without final marks, the measure used to determine programming aptitude is missing.</td>
</tr>
<tr>
<td>No comparisons of data.</td>
<td>Data compared to D&amp;B’s.</td>
<td>D&amp;B was not a replication.</td>
</tr>
<tr>
<td>Correlation examined between consistent mental models and passing/failing a course [5].</td>
<td>Correlation examined between consistency and being above/below the median.</td>
<td>The results from using a pass/fail scale cannot be used to compare other results easily.</td>
</tr>
<tr>
<td>Blank and inconsistent participants were combined during analysis [5].</td>
<td>Blank participants were not included in analysis.</td>
<td>No rationale given in D&amp;B for combining blank and inconsistent groups. We found this to be flawed. We did not have sufficient blank participants to perform a separate analysis.</td>
</tr>
</tbody>
</table>