
Avoiding “It’s *JUST* a Replication”

Bonnie E. John

IBM T. J. Watson Research Center
1101 Kitchawan Rd
Yorktown Heights, NY 10598 USA
bejohn@us.ibm.com

Abstract

This position paper explores my experiences getting replication studies accepted at the CHI conference over the past 30 years. These experiences lead to my hypothesis that CHI reviewers and program committee members at all levels need education and technology support to understand and appropriately consider replication studies for publication at CHI. I propose a draconian “zeroth iteration” on a design for extensions to the Precision Conference System to spur discussion about how we can design our values into our processes.

Author Keywords

Experimental design, replication.

ACM Classification Keywords

H5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous.

General Terms

Human Factors

Introduction

Replication has been at the heart of science for as long as the scientific method has existed; sometimes it feels as though I have been fighting for the value of replication at CHI almost as long. As an engineer by training and inclination, replication is of even more importance for the practice of UI design, in my view, because practitioners can (and should) only trust

Presented at RepliCHI2013. Copyright © 2013 for the individual papers by the papers’ authors. Copying permitted only for private and academic purposes. This volume is published and copyrighted by its editors.

results from science when the results have been replicated at several different research groups (i.e., direct replication) and the boundaries of applicability have been thoroughly explored through replicate+extend studies. I cannot count the number of times I have heard "Reject; it's JUST another Fitts's Law study" or "Reject; it's JUST another GOMS study" at program committee meetings in our field. When present, I have sometimes been able to rescue these contributions to our field's science base. I can only imagine how many such papers were rejected when I, or like-minded researchers, were not present and how many potentially-contributing authors have been discouraged by such "JUST a replication" reviews. This position paper is a proposal of how to avoid "It's JUST a replication" in the absence of dogmatic senior researchers like me.

Hypotheses about the problem

It is my experience that some sorts of replication are more acceptable to reviewers and program committees than others. The most acceptable seem to be those that replicate only a method, e.g., Baskin and John [1] used the same method of achieving extremely skilled task execution performance as did Card, Moran and Newell [2]. Using the same method to study performance on a GUI CAD system [1] and a command-line text editor [2] was not criticized by reviewers, seemingly because the tasks were sufficiently different. My hypothesis is that method replication is not a problem in HCI research publication, so much so that it might not even be recognized as a type of replication.

However, I know of replicate and extend papers falling (or being pushed) into the *JUST-a-replication* barrel

when they vary any one of the myriad other variables in a study.

Extending the participants to a new user group.

For example, a study I cannot name for confidentiality purposes was rejected when it replicated an educational treatment using participants who were different from the previously published work: they were at a lesser-known school, they were in a different major and therefore could be assumed to be less motivated to do well on a topic, and were given less direct access to expert support in doing the experimental support. The fact that these participants performed as well as the majors at a top-of-the-line school studying under the inventor of the educational treatment is a replication worth printing because it gives hope that the educational treatment will scale beyond the reach of its inventor.

Similarly, a paper that was rescued from *JUST-a-replication*, but which I will not name to maintain confidentiality, described a well-known HCI method being used by practitioners far outside the HCI field, having picked up the technique from the HCI literature and made profitable use of it, verified with empirical data. That any of our methods can be of use to people without our help is a result worth publishing because it also shows that the beneficial impact of our field can extend beyond the reach of our limited number of researchers.

Extending the measures in the study to cover new questions

Again, in a rejected paper I cannot reveal, a replication was done that included additional survey data that explored *why* some behavior was observed in both the

original and replication studies. The survey instrument was new, the data was new, and, to me, the insight it revealed was new, but this was rejected as *JUST-a-replication*. Thus, there seems to be a disagreement in our community about how much extension constitutes a publishable extension. In my opinion, the replication itself was valuable and the extension was icing on the cake, but that was not the opinion of the reviewers. Differences of opinion about what does and does not constitute a publishable contribution are not uncommon, and in fact should be encouraged, but the reviews *did not even acknowledge that there was any extension at all*, causing me to hypothesize that the definition of replicate+extend is not well assimilated into our review community.

Direct replication to increase statistical power so that new questions can be answered

Tired of not being able to give details of the papers I have discussed above, I offer my own rejected CHI paper to make a point about direct replication [4]. We had done a study with only six participants per condition and the effect was so strong that it attained statistical significance on some coarse measures and was published at the IEEE's International Conference on Software Engineering [3]. The coarse measures did not help us understand why the participants performed better on some conditions than others and did not distinguish between two conditions that had important implications for the practical use of the technique we were investigating. Therefore, we did a direct replication of the previous study, justified combining the data, and were able to tease out several new insights given the increased power of the combined study. We thought the results were a significant contribution beyond the initial study, and in fact, these

results are the only ones that excite software engineering audiences when I talk about them (SEs are the target "users" of these research results).

Whether you agree that the results are exciting enough to publish is immaterial to the reviews we received – "Reject; it's JUST a replication" without comment on the new analyses and results. This leads me to the hypothesis that new analyses are not sufficiently valued or understood by our reviewing community to warrant comment. The replication "surface structure" is enough to push a paper into the *JUST-a-replication* barrel.

And interesting point about the interaction of replication and anonymous reviewing was brought out by this paper as well. This was in the era of CHI's strict rules about anonymization, so we wrote about ourselves in the third person, as instructed. A reviewer seemed to think that using "Golden et. al's" materials was somehow cheating or lazy and criticized us for not creating our own materials. Again, this leads to the hypothesis that our reviewing community is in need of education about the process of a good replication (i.e., NOT making your own materials) and highlights a potential confound between anonymity and replication. Might the paper have been less harshly reviewed if the reader had known that we did the original study, i.e., we did do the hard work of creating the materials and were not cheating or lazy?

A proposed approach to a solution

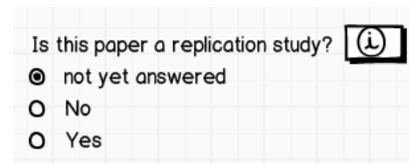
As explained above, my experiences lead me to the hypothesis that if our community is to embrace replication and publish good ones, reviewers need to be educated about what makes a good replication and its value to the field.

It is not sufficient to instruct Associate Chairs (ACs) and Sub-committee Chairs (SCs) as was done at the Program Committee meeting for CHI2013, because reviewer scores push replications down in the rankings and we cannot depend on human memory in the heat of PC debates to raise such papers to the level of discussion.

Therefore, I propose that we build our values into submission and reviewing software (Precision Conference System, PCS), to be a "job aid" to authors, reviewers, ACs and SCs, delivering education at the time it is needed. Below I present "iteration 0" of a design for these extensions to PCS.

Job aid for authors:

Present a required radio button for authors at submission time. Include an information button next to the question that leads to information about what a replication study is and what the criteria for reviewing are for a replication study.



Is this paper a replication study? ⓘ

not yet answered

No

Yes

It is possible that we would want to ask for the type of replication (direct replication, replicate+extend, or conceptual replication), but that may be introducing too much complexity in the first iteration.

Job aid for reviewers

If the author has declared the paper to be a replication study, then the review form shown to reviewers

changes to include specific required fields that apply to replication studies. Include an information button next to every field so the reviewer can get information about acceptable replication processes and the general value of replication at the time of filling out the review. Depending on how much we believe our target users need the education, we may consider presenting this information in a modal dialog box when field is first clicked by a reviewer with a button that dismisses the dialog box and a checkbox "do not show me this again" appearing after a reasonable amount of time needed to read the text in the box.

Reviewers should be able to identify themselves to PCS as being skilled in assessing replications and interested in doing so.

Job aid for Associate Chairs (ACs)

If the author has declared the paper to be a replication, this is indicated to the AC at paper-assignment time, so the AC is aware that reviewers skilled in experimental design and analysis should be recruited. Such reviewers may be self-identified in PCS, as above. We may also consider allowing ACs and SCs to identify especially skilled replication reviewers in PCS, like we currently acknowledge excellent reviews.

At review time, the AC's meta-review form also changes to include required fields that specifically address issues with replication, with information buttons.

PCS could also automatically mark this paper "to be discussed at the PC meeting". Depending on how aggressive the CHI conference wants to be that year for

considering replication papers, this status may or may not be changed by the AC.

Job aid for Subcommittee Chairs (SCs)

If the author has declared the paper to be a replication, this is indicated to the SC at the time that papers are assigned to ACs, so the SC can assign an AC skilled in assessing replication. When recruiting ACs for a subcommittee likely to get replication submissions, the SCs might be asked to identify one or two ACs who are skilled in assessing replications, which will get the SCs thinking about this necessary skill when they can do something about it instead of when replication studies arrive.

At the PC meeting, the SC's view should highlight the papers that were identified by their authors as being replication studies, so the SC can query the AC about them during the meeting. Even if PCS allows the AC to change the status of the paper to "do not discuss" it would contribute to the education of all ACs if a sentence or two were said at the PC meeting about why this replication paper was not being discussed.

Conclusion

The zeroth iteration on changes to PCS proposed above are purposely draconian to start discussion of how our

conference reviewing technology can support our value system surrounding replication studies. I believe the need is there, let's put our UI design skills and our SIG's money where our values are.

Acknowledgements

This research was supported by in part by IBM. The views and conclusions in this paper are those of the authors and should not be interpreted as representing the official policies, either expressed or implied of IBM.

References

- [1] Baskin, J. D. & John, B. E. (1998) Comparison of GOMS analysis methods. *Proceedings Companion of CHI, 1998* (Los Angeles CA, April 18-23, 1998) ACM, New York. Pp. 261-262.
- [2] Card, S. K., Moran, T. P., Newell, A.: *The Psychology of Human-Computer Interaction*. Lawrence Erlbaum Associates, Hillsdale, NJ (1983)
- [3] Golden, E., John, B.E., and L. Bass. (2006) The value of a usability-supporting architectural pattern in software architecture design: A controlled experiment. *Proceedings of the 27th International Conference on Software Engineering*, May, 2005, St. Louis, MO.
- [4] Golden, E., John, B.E., and L. Bass. (2007) Helping software developers achieve usability. Unpublished replication study.