

RepliCHI 2013

27th-28th April 2013 – Paris, France

Proceedings of the CHI2013 Workshop on the Replication of HCI Research

Executive Summary

RepliCHI2013 was the first workshop-style event in the developing RepliCHI agenda, and thus the first to accept papers about replications as submissions from the HCI community. Three position papers were accepted, including positions that were both for and against the view that replication of research is important in our field; in particular these position papers helped to both frame and, for the first time, explore the limits of the agenda. Ten Experience Reports then helped discuss the very nature of replication in the HCI community through grounded examples and real events. To support discussion, the workshop welcomed five “original authors” of work that was replicated in our submissions, which enriched the insights gained particularly when replicating other peoples work, and trying to be replicable when publishing. For an outcome, the workshop focused on further refining a) our understanding of HCI replication practices, and b) the nature of the RepliCHI event for subsequent years.

Organised by

Max L. Wilson

University of Nottingham, UK
max.wilson@nottingham.ac.uk

David Coyle

University of Bristol, UK
David.Coyle@bristol.ac.uk

Ed H. Chi

Google, Inc.
chi@acm.org

Paul Resnick

University of Michigan, USA
presnick@umich.edu

Position Papers

- Page 3 - Avoid “It’s JUST a Replication”**
Bonnie E. John
- Page 8 - Is replication important for HCI?**
Christian Greiffenhagen & Stuart Reeves
- Page 14 - RepliPRI: Challenges in Replicating Studies of Online Privacy**
Sameer Patil

Experience Reports – Presented in Tandem with Original Authors

- Page 19 - Replicating an International Survey on User Experience: Challenges, Successes and Limitations**
Carine Lallemand, Vincent Koenig & Guillaume Gronier
- Page 24 - Replicating and Extending Research on Relations between Visual Aesthetics and Usability**
Noam Tractinsky
- Page 29 - Replicating and Extending a Facebook Uses & Gratifications Study: Five Years Later**
Tasos Spiliotopoulos & Ian Oakley
- Page 34 - NewsCube Replication: Experience Report**
Sidharth Chhabra & Paul Resnick
- Page 39 - Teaching HCI Methods: Replicating a Study of Collaborative Search**
Max L. Wilson

Experience Reports

- Page 44 - Do lab effects transfer into the real-world? And should we care?**
Petr Slovak, Paul Tennent & Geraldine Fitzpatrick
- Page 49 - Re-testing the Perception of Social Annotations in Web Search**
Jennifer Fernquist & Ed H. Chi
- Page 53 - Challenges of Replicating Empirical Studies with Children in HCI**
Quincy Brown, Lisa Anthony, Robin Brewer, Germaine Irwin, Jaye Nias & Berthel Tate
- Page 58 - Replicating Residential Sustainability Study in Urban India**
Mohit Jain, Yedendra B. Shrinivasan & Tawanna Dillahunt
- Page 63 - Replicating and Applying a Neuro-Cognitive Experimental Technique in HCI Research**
David Coyle
- Page 67 - Replicating Two TelePresence Camera Depth-of-Field Settings in One User Experience Study**
Jennifer Lee Carlson, Mike Paget & Tim McCollum

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

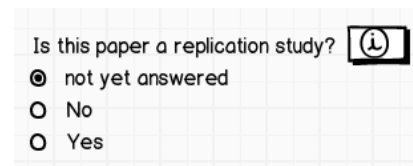
As explained above, my experiences lead me to the hypothesis that if our community is to embrace replication and publish good ones, reviewers need to be educated about what makes a good replication and its value to the field.

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.

Is replication important for HCI?

Christian Greiffenhagen

Loughborough University
c.greiffenhagen@lboro.ac.uk

Stuart Reeves

University of Nottingham
stuart@tropic.org.uk

Abstract

Replication is emerging as a key concern within subsections of the HCI community. In this paper, we explore the relevance of science and technology studies (STS), which has addressed replication in various ways. Informed by this literature, we examine HCI's current relationship to replication and provide a set of recommendations and points of clarification that a replication agenda in HCI should concern itself with.

Author Keywords

Replication; psychology; science and technology studies; philosophy of science.

ACM Classification Keywords

H.5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous.

Introduction

Replication is emerging as a concern within subsections of the HCI community. A key motivation for this is a feeling that HCI emphasises *novelty* over *consolidation* of research; consolidation that can be achieved *via replication*. In response, we advocate the relevance to HCI of understandings of 'replication' emerging from the philosophy and sociology of science and technology. This paper highlights a collection of rejoinders to the ways in which this programme for replication is currently conceptualised within HCI. In doing so we intend to help the development of an endogenous

understanding of replication as a practice that can be a) motivated, b) mature and c) fit for the purposes of HCI.

Replication: Lessons from STS

We believe that debate on replication in HCI can be enriched by STS and philosophy and sociology of science. In this section we review some of the findings of this literature and their pertinence to HCI.

One of the motivations for replication within HCI is the wish to make HCI more scientific by modelling HCI on other sciences (e.g., "psychology, physics and medicine" [11]). While there is nothing problematic in asking for a field to involve more replication, to frame this in terms of making it more 'scientific' is possibly based on a mythical view of 'good science' of which "[r]eplication of research is a cornerstone" [11]. This view suggests this 'science' may be a homogenous practice, possibly even based around a particular method, 'the' scientific method. It also tends to think about replication more from the perspective of the *philosophy* of science, rather than the *practice* of different sciences.

In contrast, philosophical and sociological studies have shown that 'science' refers to a fragile structure of multiple disciplines and multiple methods linked by 'family resemblances' only [9, 3]. Not all empirical sciences work with experiments, and the role of experiments may differ between different fields.

Complicating this picture is the separation between these varied and autochthonous scientific practices and their rendering into literature. Scientific literature is written in such a way as to promise replicability, emerging from Boyle's attempts to create scientific records that were

publicly accountable and would let 'anyone' replicate experimental practices [10]. However, the nature of instructions is such that they are always incomplete [4], thus scientific instructions must be 'filled in' by competent members of the target scientific community in order to enact them as replications. This is one of the reasons why Medawar characterised the scientific paper, somewhat misleadingly, as a 'fraud' [8].

STS reports an alternative view on the nature of replication in the natural sciences to the surface view of scientific replication where scientific articles (in particular: their 'method sections') provide an adequate instruction manual for replication work. Specifically it problematises the notion of a 'decisive experiment' or by extension a 'decisive replication'. At the heart of this problem is what Collins calls the "experimenter's regress" [1], that is, a circular relation between experimental findings and the instruments used to produce them. Reliable experimental findings themselves rely upon reliable instruments and vice versa. As a result, a key difficulty of replication and the experimenter's regress is that, particularly for contested science, there is not necessarily any standard for what is to be considered a valid replication. This raises a principle problem, since it is not clear whether a 'failed' replication is due to a problem with the original experiment or the subsequent replication ("it is often hard to tell whether an inability to replicate a result is due to a group's failings or a flaw in the original paper" [5, p. 345]).

Further to this, when we consider the track record of replication in the natural sciences, STS literature argues that replication in the (natural) sciences employs replication for specific, highly motivated and reasoned

ends. Thus we find a marked absence of large amounts of replication in the sciences unless we focus on particular issues [1, pp. 210-211]. For instance, Collins' tracing of the construction of gravitational wave detectors during the 1970s reveals the relevance of replication as an activity for working through what was a contested, controversial domain [2]. In short, 'doing replication' is not always seen as a fundamental prerequisite for valid scientific practice, since a vast number of results go unreplicated: instead it emerges as the result of pragmatic action for specific contested cases.

In summary, then, our cursory examination of STS and its related literature highlights that: a) there is no singular form of science or scientific method upon which to model; b) there is no 'algorithmic' method for replicating directly from scientific literature (indeed, this is not its purpose); c) 'absolute' security of results is problematic in light of the experimenter's regress; and d) sciences often do not involve replication as a 'matter of course', it being difficult and of little value unless motivated (typically via contestation of results).

Replication within HCI

This issue of replication has become a centre of discussion within HCI. In light of STS's view on replication, we seek to ask what is at stake in this discussion. Why replicate? Or: What are the (different) *aims and motivations* for replication?

Within HCI, it has been acknowledged that there is not just one kind of replication. For example, Wilson et al. distinguish between four forms: "direct replication" ("driven by the aspirations of strong science"), "conceptual replication" (replication via "alternative methods"), "replicate & extend" (building on prior

studies incrementally) and "applied case studies" (replication through application of prior work) [11].

Nuancing this view, we want to start with introducing two different kinds of distinction to help us to think about replication.

The first distinction is between what we characterise as *textbook replication* and *frontier replication*. By 'textbook replication' we refer to replications of well-known studies that are conducted from HCI textbooks, typically as part of undergraduate or graduate education. For instance, these could be replications of well-known usability studies. We distinguish this from 'frontier replication' by which we mean replications of 'ongoing' or 'recent' studies. We see these forms as conceptually and practically incommensurate, as opposed to integral facets (e.g., see position in [11] on "Benefits of Replication"). Thus, while the primary aim and motivation of textbook replications is *learning*, the point of frontier replication is often a form of 'checking' (which may even be done during the review process). As such we argue that the activities at this 'frontier' becomes the main issue for replication rather than what is happening 'in the textbook'.

A second distinction has to do with what may be *replicable* and what is actually *replicated*, in which the aims for each are quite different. 'Being replicated' concerns the 'factual' question of whether a particular study has, actually, been replicated by other researchers or not. We say 'factual' since subsequent studies may or may not be seen as valid replications, as in Collins' study of gravitational wave detectors [2]. We also note again that a lack of actual replications may be related to matters such as experiments being too costly, too time

consuming or lacking in providing the experimenter any obvious credit.

In contrast 'being replicable' is motivated by the 'in principle' possibility of some other researcher being able to replicate an empirical study. This is often cited as one of the differences between 'quantitative' and 'qualitative' methods (very problematic descriptions themselves), where the former supposedly produce results that *could* be replicated (again, 'in principle'), while the latter are not. For instance, ethnographic research is often said to be too reliant on the 'subjective' insights of the ethnographer, resulting in non-generic and non-replicable findings.

What's at stake in this distinction? We would argue that the issue of 'being replicable' concerns a foundational question, in particular, whether HCI is a science and its preferences for particular methods over others. These questions are not new: psychology—which has strongly informed HCI's development—has repeatedly foregrounded replication as an explicit agenda, such as in response to perceived experimental biases (e.g., being too 'WEIRD' [6]), as well as intentional and unintentional misconduct [12]. In this sense, 'being replicated' is probably more common in psychology than many other sciences because of this explicit concern (now displayed in HCI) for the lack of actual replicated studies (or those 'seen as' validly replicated).

Psychology's own debates around its status as a science are also consonant with these foundational concerns of 'being replicable', and in the replication agenda we see HCI grasping towards key epistemological themes which arise in the natural sciences: alongside 'observation', 'measurement',

'description' and 'reasoning' is, of course, 'replication'. If we take HCI as a scientific endeavour (e.g., [11]) then it follows that its concern for replication would thus be informed by this particular picture of 'normal science'; or 'doing what scientists do'. However, this assumes coherence of 'science' as monolithic practice as well as mythologising that practice.

In contrast, 'being replicated' is a more pragmatic question, which concerns what we can *learn* from replications and, for example, whether it would be worthwhile to publish more papers based on replication.

In order to focus the discussion of replication in HCI, it would be very helpful if one could gather more examples from different disciplines, from biology to physics, to see whether and how replications are valued in these. Thus we hazard a conjecture: that replication enjoys a special status within psychology (and the debate of replication in HCI is thus a reflection of the influence of psychology, rather than, say, biology, in HCI). But why might that be?

One issue is with the *scale* of the question to be answered through experiment. Some sciences tackle very detailed and small questions through extremely detailed experiments. In other words, there exist a very tight relationship between the data gathered through the experiments and the derived conclusions. Other sciences (e.g., social science) tackle bigger questions and consequently involve a looser relation between data and conclusion.

We would argue that there is a 'scale' tension in psychology—and thus HCI—between tackling 'big' and 'minute' questions, questions that can, or can't be

settled through experiments. One possible reason for more replication in psychology is that studies can be questioned more (i.e., findings are more open to interpretation).

Discussion

We have raised some broad issues in the relationship between replication and HCI, and informed this debate through recourse to existing work in STS that has explored replication in the natural sciences.

Firstly we argue for the importance of the increased consultation of literatures normally foreign to HCI such as that of STS. This is particularly the case for situations where knowledge within the field is out of step with more recent advances in understandings of scientific knowledge. For instance, our discussions on replication (and science) within HCI are largely Popperian or pre-Popperian in form, such as appeals to ideals such as falsificationism. While we would not argue against such ideals, we contend that understanding benefits from expansion, thus as well as citing Collins, we might also refer to developments by Kuhn, Feyerabend or Lynch that, for instance, encapsulate empirical investigations into practical mundane scientific action [7].

A fundamental question for the desire for replication in HCI is that of the motivation to perform replication in the first place. We need to ask ourselves *why we might bother* with replication in the first place and whether there is any value gained from pursuing a replication agenda as a distinctive activity within HCI (which is the position of the workshop call [11]). As we have seen from STS literature, if we feel the need to derive HCI's programme from the methods and epistemological

topics of the natural sciences (e.g., via psychology), then we must do so knowingly in light of findings from STS. Thus we argue for different understandings of replication: a) as an unstable and negotiated practice; b) as a highly motivated activity rather than as an end of itself; and c) as playing an important role in the resolution of scientific controversies. Moving forwards we would draw attention to the judicious motivated application of replication—and the need for 'just why' and 'just how' it is to be pursued. So, we must be clear about the purposes and motivations of any given replication beyond abstractly "validating and understanding contributions" [11].

Finally, we have argued that a mythological view of science tends to be implicit in HCI regarding its status as scientific. This leads us to question the value in positioning HCI as a scientific endeavour. Thus we recommend that it would be helpful to separate the 'foundational' question (whether HCI is a science) from the above 'pragmatic' question (about the specific benefits of replication for HCI).

Acknowledgements

This work is supported by Horizon Digital Economy Research, RCUK grant EP/G065802/1.

References

- [1] Collins, H. M., *Changing Order: Replication and Induction in Scientific Practice*, Beverley Hills & London: Sage, 1985.
- [2] Collins, H. M. The seven sexes: A study in the sociology of a phenomenon, or the replication of experiments in physics. *Sociology*, 9(2):205-224, 1975.
- [3] Dupre, J. The disunity of science. *Mind* 112, 321-346, 1983.

- [4] Garfinkel, H. *Studies in Ethnomethodology*. Prentice-Hall, 1967.
- [5] Giles, J. The trouble with replication. *Nature*, 442:344-347, July 2006.
- [6] Henrich, J., Heine, S. J. and Norenzayan, A. The weirdest people in the world? *Behavioral and Brain Sciences*, 33, pp. 61-83, 2010.
- [7] Lynch, M. *Scientific Practice and Ordinary Action*. Cambridge University Press, 1993.
- [8] Medawar, P. B. Is the scientific paper a fraud? *The Listener*, 70 (12 September): 377-378, 1963.
- [9] Putnam, H. The idea of science. *Midwest Studies In Philosophy*, 15(1):57-64, 1990.
- [10] Shapin, S. and Schaffer, S. *Leviathan and the Air-Pump: Hobbes, Boyle, and the Experimental Life*. Princeton University Press, 1989.
- [11] Wilson, M. L., Resnick, P., Coyle, D. and Chi, E. H. RepliCHI—The Workshop. In *CHI '13 Extended Abstracts (CHI EA '13)*. ACM, New York, NY, USA, 2013.
- [12] Yong, E., Replication studies: Bad copy. *Nature*, 485(7398):298-300, 2012.

RepliPRI: Challenges in Replicating Studies of Online Privacy

Sameer Patil

Helsinki Institute for
Information Technology HIIT
Aalto University
Aalto 00076, Finland
sameer.patil@hiit.fi

Abstract

Replication of prior results has recently attracted attention and interest from the CHI community. This paper focuses on the challenges and issues faced in carrying out meaningful and valid replications of HCI studies. I attribute these challenges to two main underlying factors: (i) a domain of inquiry that simultaneously covers people, social systems, and technology; and (ii) deficiencies in result reporting and data archiving. Using examples from investigations of online privacy, I outline how these challenges manifest themselves in HCI studies. Longitudinal approaches, international collaboration, and sharing of study instruments could help address these challenges.

Author Keywords

Replication, Privacy, Cultural differences

ACM Classification Keywords

H.1.2 [User/Machine Systems]: Human factors.

General Terms

Human Factors, Security

Introduction

Replication of prior results has recently attracted attention and interest from the CHI community. The

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.

resulting discussions tackle replication from two important perspectives: higher level epistemological debate on the place and merits of replication in the scientific (publishing) enterprise and the lower-level practical considerations for replicating previous studies from the literature. Growing interest in RepliCHI suggests increasing recognition for the value of replicating prior studies. I hope and anticipate that this trend will foster continued community discussion on how to justify, appreciate, and reward replication as a valuable scientific pursuit. Therefore, in this paper I focus on the latter aspect, viz., challenges and issues faced in carrying out meaningful and valid replications of HCI studies.

I attribute these challenges to two main factors:

1. Domain of inquiry: A large proportion of HCI studies tackle research problems where results typically exhibit simultaneous and interacting influence of individuals, social systems, and technology. Each of these three factors changes at drastically different rates and magnitudes. For instance, technology used in a study may become obsolete within months or a couple of years, while physical and cognitive capabilities of adults change at much slower rates (and the magnitude of the change is often comparatively small and predictable). These differences in the evolution trajectories of humans, cultures, and technology make it difficult to replicate studies at a later time and to determine and attribute causes behind differences in results, if any.
2. Insufficient and/or incomplete reporting: Typically the only resource available for replicating a study

is the publication describing the results of the study. Unfortunately, due to page limits and other editorial reasons, publications often do not include all information — about methods and/or data — necessary for carrying out the study the way it was originally conducted. For instance, instead of including the entire questionnaire instrument, the publication may include only those questionnaire items that led to statistically significant results. Similarly, results may be presented in the aggregate or as percentages, making it difficult to replicate analyses that require details of individual data points.

In the following section, I outline how I have found these challenges to manifest themselves in investigation of user preferences and practices regarding online privacy. I conclude with some thoughts on addressing the challenges.

Replicating Studies of Online Privacy

When thinking about and carrying out replications of research related to privacy, I have encountered several practical challenges:

Privacy is a nuanced and complex issue affected by individual characteristics, context of operation, and the technology under consideration. For instance, individuals have been classified into different groups based on their inherent level of privacy concern [7], and privacy concerns have been shown to exhibit cultural variation [3]. People's mental models and understanding of the underlying technology also affects their preferences and practices regarding privacy [4]. This implies that even when considering the *same* technology, replication conducted at a later time ought

to take into account the impact of learning effects on privacy issues. Replications may also encounter the selection-maturation threat to validity owing to major external events that occur after the original study, such as news coverage of privacy breaches. Such events affect the population's overall understanding and awareness of privacy issues, thereby potentially affecting the results of replications of studies that were originally conducted prior to these event(s).

The majority of attention in replication has been devoted to replication at a different (later) *time*. In the case of privacy, however, it is equally important to consider replication across different *cultures*. For example, we administered a questionnaire simultaneously in the US and India, enabling us to draw interesting and surprising observations from comparison across cultures [5]. Our results confirmed earlier findings regarding low levels of consumer privacy concerns in India. Surprisingly, by examining interpersonal privacy separately from consumer privacy, we found that interpersonal privacy concerns in India were not only higher than consumer privacy concerns but also higher than interpersonal privacy concerns in the US. Our study considered culture at the broad level of national cultures. However, it should be noted that for replication purposes "culture" could be construed to connote any large groups with shared characteristics and/or values, such as students, engineers, mothers, liberals, etc. Moreover, if replication across cultures is conducted at a time later than the original study, then learning effects and maturation threats need to be taken into account (as discussed above).

In theory, replication with a different cultural sample is

a simple case of re-running the study with subjects drawn from a different culture, with translation of instruments and study materials, if necessary. In practice, however, cultural differences pose several hurdles. For instance, the same word or term may be interpreted differently leading to the same question being answered differently. For example, we found that the term "cubicle" was understood differently in the US and India owing to differences in office layouts and density. This difference was one of the factors crucial for understanding the differences in results between the US and India [5]. In other studies, I discovered that the demographic question about ethnicity, which is commonly asked in the US (and even mandated for NSF-sponsored studies), was considered potentially offensive and confusing in Europe. Differences in lifestyle and beliefs can also affect whether questions and tasks from one study can yield valid results, or even make sense, when replicated in a different cultural context. For instance, some privacy studies have asked Western respondents about premarital sex, sexual practices, extramarital affairs, and number of sexual partners (e.g., [1]). Such questions are unlikely to produce meaningful results in cultures where such practices are uncommon and/or forbidden. Resolving this issue can be complicated when such culturally-specific questions comprise parts of standard scales; using the scale without modifications will not yield meaningful results and dropping and/or modifying items in the scale risks affecting the validity of comparison across studies. Finally, it is also necessary to consider whether results across cultures are affected by differences in sampling techniques and sample characteristics. For instance, although our comparison of the US and India was limited to software professionals, the mean and median ages of the Indian

participants were lower than those of the US participants.

We found that understanding privacy-related cultural nuance often requires insights derived from qualitative methods (such as interviews, focus groups, field visits, etc.) and/or insider knowledge of the culture and its practices [6]. Currently the CHI community is focused mostly on replication of studies that employ quantitative methods, such as experiments, questionnaires, or usability evaluations. Complementing quantitative replications with qualitative insights has potential to broaden the scope of these replication endeavors. Toward this end, it may also be fruitful to tackle whether and how qualitative studies could be effectively replicated.

Discussion and Conclusion

The previous section utilized examples from investigations of online privacy attitudes and behaviors to illustrate some of the challenges and issues in replicating HCI studies. Online privacy cuts across the individual, the social, and the technical, in much the same way as many studies in HCI do. Therefore, I believe that many, if not all, of these concerns are also likely to arise in HCI investigations of other topics.

The RepliCHI workshop is an important milestone toward developing a comprehensive compilation and understanding of various challenges involved in the replication of HCI studies. Moving forward, it is necessary to apply this knowledge and insight for constructing best practices to follow and pitfalls to avoid. Toward this end, I offer suggestions that address the two important considerations outlined in the Introduction, viz., (i) domain of inquiry that

simultaneously covers individuals, social systems, and technology; and (ii) result reporting and data archiving.

The second of these, in particular, could be easily addressed by requiring inclusion of full instruments and study protocols as appendices¹. Similarly, authors of accepted papers could be asked, or even required, to upload the raw data after taking steps necessary to protect participant anonymity. In this regard, ACM, IEEE, NSF, and other prominent HCI funding and sponsoring organizations can follow the lead of the NIH, which mandates raw data availability. In a similar vein, an open source inspired approach could encourage authors to release the source code of systems and scripts used for conducting studies and carrying out analyses. An open question regarding data and code sharing is how to deal with commercialization and intellectual property issues (especially when corporate entities are involved in conducting the study)².

One approach for addressing the issue of intersection of people and technology is to encourage longitudinal investigations carried out at regular intervals over several years. Depending on the details and logistics of the study, a longitudinal investigation could utilize the *same* participants or different participants with the same sampling method and sample characteristics. The former approach can help examine the impact of changes in individual characteristics, evolution in lifestyles, and effects of learning. The latter approach can help illuminate the impact of changes in

¹This also provides the additional benefit of addressing one of the most common comments raised in peer reviews — lack of methodological detail.

²Data used by studies conducted by corporations was a hotly debated topic at the WWW 2012 conference [2].

technology. For replications across cultures, however, it is perhaps best to target simultaneous study deployment. Fostering international collaborations and/or leveraging international students to gain cultural knowledge and access could help in this regard.

Requiring a replication component in Bachelor's and Master's theses could provide a starting point for repeating studies from the literature, simultaneously serving a valuable pedagogical purpose by training the next generation. Further, conferences and journals could explicitly solicit replications of specific studies. Special conference sessions or journal sections could be devoted solely to replication studies. Discussions and follow-up activities from the RepliCHI workshop could lead the way toward legitimizing and promoting replication as a valuable scientific pursuit within HCI.

Acknowledgments

I thank Mihir Mahajan and John McCurley for editorial comments.

References

- [1] Grossklags, J., and Acquisti, A. When 25 cents is too much: An experiment on willingness-to-sell and willingness-to-protect personal information. In *Workshop on the Economics of Information Security (WEIS)* (2007).
- [2] Markoff, J. Taves of personal data, forbidden to

researchers.

<http://www.nytimes.com/2012/05/22/science/big-data-taves-stay-forbidden-to-social-scientists.html>, May 2012.

- [3] Milberg, S., Burke, S., Smith, H., and Kallman, E. Values, personal information privacy, and regulatory approaches. *Communications of the ACM* 38, 12 (1995), 65–74.
- [4] Patil, S., and Kobsa, A. Uncovering privacy attitudes and practices in Instant Messaging. In *Proceedings of the 2005 International ACM SIGGROUP Conference on Supporting Group Work*, GROUP '05, ACM (New York, NY, USA, 2005), 109–112.
- [5] Patil, S., Kobsa, A., John, A., and Seligmann, D. Comparing privacy attitudes of knowledge workers in the U.S. and India. In *Proceedings of the 3rd International Conference on Intercultural Collaboration*, ICIC '10, ACM (New York, NY, USA, 2010), 141–150.
- [6] Patil, S., Kobsa, A., John, A., and Seligmann, D. Methodological reflections on a field study of a globally distributed software project. *Information and Software Technology* 53, 9 (2011), 969–980.
- [7] Taylor, H. Most people are “privacy pragmatists” who, while concerned about privacy, will sometimes trade it off for other benefits. *The Harris Poll 17* (2003), 19.

Replicating an International Survey on User Experience: Challenges, Successes and Limitations

Carine Lallemand

Public Research Centre Henri Tudor
29 avenue John F. Kennedy
L-1855 Luxembourg
Carine.Lallemand@tudor.lu

Vincent Koenig

EMACS Research Unit &
Interdisciplinary Centre for
Security, Reliability and Trust
University of Luxembourg
Route de Diekirch
Walferdange, L-7220
Luxembourg
Vincent.koenig@uni.lu

Guillaume Gronier

Public Research Centre Henri Tudor
29 avenue John F. Kennedy
L-1855 Luxembourg
Guillaume.Gronier@tudor.lu

Abstract

In order to study how the notion of User Experience (UX) evolved over the last few years, an international survey originally conducted in 2008 by Law et al. [1] has been replicated. Its main goal was to get some insights on the points of view from practitioners on the notion of UX. After having slightly adapted the initial (English) survey and having translated it into French and German, more than 758 valid answers have been collected from all over the world. This experience report aims at illustrating some of the challenges involved in the replication of such a study as well as successes and limitations.

Author Keywords

User Experience; HCI Research; Replication; Survey; Experience Report

ACM Classification Keywords

H.5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous.

General Terms

Human Factors; Design; Measurement

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.

Introduction: a Tale of Two Studies

Some concepts in the field of HCI are widely spread and used by practitioners even if a lack of empirical research prevents the true understanding of their meaning and impacts [3, 1, 2]. This is the case for User Experience (UX). Despite many attempts to understand, define and scope UX, it is still not clear whether a consensus has been reached on this concept or not. In a willingness to address the complexity of the UX concept, to contribute to its further development and consolidation, we decided to replicate a previous survey entitled "Understanding, scoping and defining UX: a survey approach" [1].

The original study has been first spread during the main conference CHI'08 before being broadcast through several communication channels. Results have been published the following year in the proceedings of CHI'09, as a 10-pages long paper. 275 answers had been collected at that time from 25 countries.

In order to adapt to our project's multicultural context and to reach a wider audience within the French-speaking community of UX practitioners, all questionnaire items have been translated, from the English master version to French and German (both languages being commonly used in Luxembourg). A back translation process has been applied to ensure the quality and validity of the process.

Rationale for a Replication

Several reasons may explain the choice to replicate this UX survey. First of all, as User Experience is still a concept in maturation, it was worth taking stock of the situation four years after the initial study in order to see a possible evolution in the representations, points

of view and practices associated to UX. Replication acts here as a way to check whether the results still apply in a different context to the original study, especially in a different temporality.

Moreover, the translation into two others languages allowed us to reach a wider audience, especially in the multicultural context in which the present work was involved. As this study constituted an exploratory step within a wider Luxemburgish project focused on UX Design, gathering additional knowledge about the French- and German-speaking practitioners' community (not well represented in the initial study) seemed crucial to us. By trying to draw an accurate picture of the current situation of UX and building on that basis, we aim at achieving the best solutions possible to design for UX.

Form of Replication

This study may be considered as a direct replication, since differences between both studies are limited to:

- A minor extension through the translation in French and German languages. The original English version was kept as default language and still represented 58.4 % of the completed surveys.
- Additional sociodemographics items aimed at better categorizing participants and acting as control variables to analyze the data.

Summary of the Methodology

Structure of the Survey

The UX questionnaire encompasses 3 sections:

- *Background*: respondents were asked to first answer 13 questions about their job and educational background, their level of familiarity with UX or the importance of UX in their actual

work. Sociodemographic information (age, gender, country of residence) was also collected.

- *UX Statements*: respondents were asked to assess their agreement level with 23 UX statements on a 5-point Likert scale.
- *UX Definitions*: Five UX definitions were presented. For each of them, participants were asked to answer the following open-ended question "What do you think of this definition?". Finally, participants were asked to choose which definition suits them best and to freely comment on the reasoning for their choice.

The main differences between the initial study and the replication lay in additional sociodemographics to better categorize respondents. The following questions have therefore been added to the initial survey: current job position, level of familiarity with the concept of UX and collaboration with people working in the field of UX.

Sampling and Dissemination of the Survey

The survey was broadcast online from February to April 2012, on multiple communication channels. As for the original study, practitioners' forums, social networks and mailing lists were the main vector of dissemination. From a total of 898 returned questionnaires, 758 valid questionnaires have been retained to compute the data.

Results

Our results mainly confirmed the original findings on the understanding of UX. Our classification of UX statements sorted by mean-agreement is very similar to the original one. Uniqueness of an experience, importance of social and cultural context, and finally temporal dynamics remained highlighted as crucial by

the respondents. Interestingly, our larger sample size allowed us to identify some patterns describing how the differences in UX perception and choices of a UX definition significantly vary with background variables. Analyses of qualitative data (open-ended questions) are still ongoing and may show differences between the replication and the original study. These questions will indeed probably allow us to identify a range of issues that may be underlined by the respondents in 2012 but were not previously conceptualized through the UX statements defined in 2008.

Challenges, Successes and Limitations of the Replication

Volatility of concepts in the field of HCI

Repeating a conceptual survey presents inherent challenges due to the relative volatility of some concepts and notions developed in HCI, but also due to the volatility of the main object of HCI. Driven towards novelty and innovation some terms used in this research field tend to emerge as popular trends and fade away quickly without having been really analyzed through the lens of empirical research. Some authors in HCI suspect that it could have been the case for UX, which is often used as an umbrella term to designate a wide range of fuzzy and dynamic concepts such as affects, hedonism or aesthetics [2]. Moreover, after 4 years of intensive use by both practitioners and researchers, it was a bit of a challenge to dare repeat such a survey aimed at defining UX - going back to the basics in a way. We had e.g. the case of a group leader on LinkedIn who refused to broadcast the study claiming that it was now useless because every good practitioner knows what UX is, even though he was unable to provide an accurate definition of UX. Fortunately, beyond this single case, the replicated

survey has been received warmly by the community, which demonstrates the need to reflect and examine the concept of UX once again, in a new temporal context. Understanding and validating previous findings seemed nevertheless highly valuable and our approach truly succeeded at analyzing the maturational process of the concept of UX.

Language and Translation of Material

When working in a non-English speaking country, replication (or even partial use of existing tools only) generally involves the translation of those tools into the native language of the users composing the target population and sample. The administration of a questionnaire in the native language of respondents allows to give them a better understanding of the items and to decrease the rate of people being excluded or who abort due to language difficulties. However, translating a survey may become very complex when dealing with conceptual topics (as it is the case here), which already involve several ambiguous items (whether intended or not by their authors) in their original version. The present study was translated into German and French. Even if a back translation process has been used to verify the reliability of the translation, it is not yet sure that concepts were understood in the same way across different languages (and maybe even across different respondents for the same language). To overcome this difficulty when computing the data, note that we also compared the level of non-understandability of the items (respondents had the option to check "I don't understand"). Being almost similar for each language and similar to the level found in the original study, the translation was considered fairly reliable.

Comparability of the results

SAMPLING AND DIFFUSION OF THE SURVEY

Replicating a research work dealing with the definition of a concept implies reaching a comparable sample both in terms of sample size and characteristics. However, how should we deal with this kind of exploratory survey that did not involve a random and representative sample? As the whole population of practitioners working in a field related to UX is not clearly defined, it was decided to simply broadcast the survey on the web. We were aware that several biases may have impacted previous results (and may also impact ours), especially the fact that only self-motivated and careful respondents would answer the questionnaire. Moreover, it was impossible to know with accuracy neither the number of people touched by the survey (probably thousands of them), nor the coverage of the target population. However, every research design choice has strengths and weaknesses. The diffusion method chosen for the original study has clearly advantages in terms of reaching a wide audience, which fulfilled the primary exploratory goal of the study and provided us with information on what kind of practitioners declare working directly or indirectly on topics related to User Experience. We succeeded in reaching an international sample larger than the original one ($n=758$ in 2012 vs. $n=275$ in 2008) but still almost equivalent in characteristics. The larger sample size had two main advantages: first it allowed detecting more subtle differences in the understanding and perceptions of the notion of UX according to background variables; second it allowed detecting societal evolution related to the field of HCI (e.g. an increase in the number of UX practitioners coming from Asia, Middle-East or Africa).

LIMITATIONS HIGHLIGHTED IN THE ORIGINAL SURVEY DESIGN

Replicating research implies repeating a study exactly the way it has been conducted the first time. Unfortunately, it is close to impossible to design studies without any limitation and thus most studies present some limitations, highlighted by the authors or not, that need to be copied for the sake of replication. While this is not intended to depreciate previous work at all, it should highlight that repeating mistakes or inaccuracies may be hard to accept as researches always strive for progress. In the case of the UX Survey, we noticed some possibilities for improvement regarding the survey design (e.g. reduction of the number of items, rephrasing of ambiguous UX statements, rotation/counterbalancing of items or reflection on open-ended questions). These improvements could have been done quite easily with a new pre-testing phase involving a few users. Although we were aware of those limitations, replication forbids any major changes in the survey design (since it may bias the results) and we had to accept this as a matter of fact. The solution we found to overcome this issue was to extend our data collection. As some data cannot be easily quantified, and as this is especially the case here when dealing with a conceptual representation of User Experience, additional in-depth interviews with practitioners were conducted in order to better understand their representations of the concept and the way they made use of it. Concomitant with the diffusion of the UX Survey, 25 interviews were conducted during the first semester 2012. A semi-directive interview guide has been created, mainly based on the principal questions included in the UX Survey [1].

Conclusion

By replicating a previous UX survey, we intended to gain further insight into the maturational process the concept of UX undergoes. Further, we aimed at validating previous findings almost taken for granted by the HCI community (e.g. uniqueness of an experience, influence of the context, or temporal dynamics of UX). Despite some challenges and difficulties to overcome, replication of such a survey appeared valuable and highly interesting for the community. Every research design has strengths and weaknesses, requiring choices to be made with regard to the research objective. Replicating a research work therefore implies both benefits from the strengths and applying the limitations of the original study.

Acknowledgements

The present project is supported by the National Research Fund, Luxembourg. The authors would like to thank the authors of the initial survey conducted in 2008 [1], especially Effie C. Law and Marc Hassenzahl, for their availability and support.

References

- [1] Law, E., Roto, V., Hassenzahl, M., Vermeeren, A. & Kort, J. (2009) Understanding, scoping and defining UX: a survey approach. Proc. CHI 2009, Boston, USA.
- [2] Law, E., & Van Schaik, P. (2010). Modelling User Experience – An agenda for Research and Practice. *Interacting with computers*, 22, 313-323.
- [3] Roto, V., Law, E., Vermeeren, A., & Hoonhout, J. (2011) User Experience White Paper: Bringing clarity to the concept of user experience. Result from Dagstuhl Seminar on Demarcating User Experience, Finland.

Replicating and Extending Research on Relations between Visual Aesthetics and Usability

Noam Tractinsky

Ben-Gurion University of the Negev
Beer Sheva
Israel
noamt@bgu.ac.il

Abstract

This paper describes a replication and extension of a study that found strong positive correlation between evaluation of a product's beauty and pre-use perceptions of its usability. The original study was conducted in Japan; its replication and extension took place in Israel. The extension involved mainly methodological improvements to the original study, which demonstrated the robustness of the original study's findings.

Author Keywords

Replication; Visual Aesthetics; Perceived usability, Cross-culture, Method bias, HCI.

ACM Classification Keywords

H.5.2. Information interfaces and presentation (e.g., HCI): User Interfaces

Introduction and Motivation

At CHI '95, I attended a session in which Masaaki Kurosu presented a short paper. The paper described an experiment, designed to find whether people's perceptions of usability (operationalized as ease-of-use) correlate with established user interface design guidelines (Kurosu and Kashimura, 1995). Kurosu and

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.

Kashimura (K&K) projected 26 different designs of ATM interfaces to groups of participants seated in a classroom. During the projection of each slide the participants rated the design in terms of its perceived usability and beauty. Evaluations of usability were then correlated with the degree to which the designs followed usability guidelines. One of the study's results, however, pertained to the relation between the participants evaluations of usability and beauty (See Figure 1). To the best of my knowledge, that study was the first in the HCI literature to provide empirical evidence regarding the relation between these two system aspects. Surprisingly, the data indicated that people's perceptions of system's aesthetics are strongly and positively correlated ($r=0.59$) with their perceptions of the system's usability.

I was surprised by K&K's findings, and thought that their study should be replicated for several reasons. First, their results ran contrary to the prevailing thought in the field of HCI. At that time beauty (or visual aesthetics) was a marginal factor in HCI research and practice. It was usually ignored; rare acknowledgments of aesthetic design were immediately followed with caveats against overemphasizing it or with a message belittling its role relative to more utilitarian aspects and objectives of interactive systems.

Second, I was willing to accept that K&K's findings may hold in the particular locale of their study – Japan – a country with a long and glorious aesthetic tradition. However, I was skeptic about the generalizability of these findings to other places. More specifically, I found it unreasonable that similar correlations would be found in my own country – Israel – which is known more for

its people's orientation to act rather than for its aesthetic tradition.

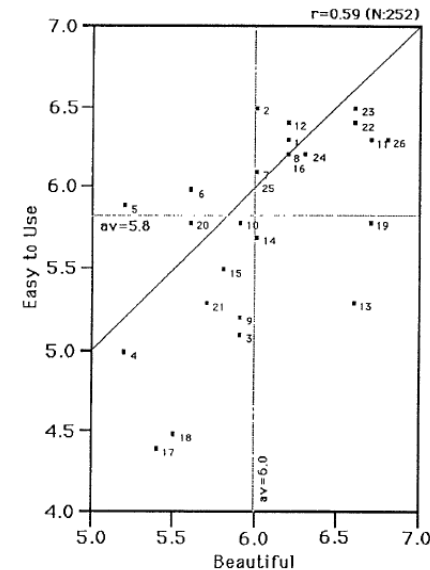


Fig.1 Correlation between two kinds of judgements for 26 layout samples.

Figure 1. Relationships between beauty and perceived usability as reported in Kurosu and Kashimura's (1995) study.

Finally, the method used in the original study was not flawless. In particular, there was a concern that the observed correlations between beauty and perceived usability were influenced, at least partially, by common method bias.

Thus, I embarked on a research project whose main objective was to demonstrate that K&K's findings were either wrong (as implied by main stream HCI

literature), or at best qualified by cultural factors (as implied by my own experience). The project, , which is described below, included replication and extension of Kurosu and Kashimura's research. Its outcomes were published at CHI '97 (Tractinsky, 1997), and are summarized below.

Replication and Extension

My research included three studies: A replication of the original study and two extensions. All three studies used the same independent and dependent variables as the original study. The stimuli (designs of ATM machines) were basically the same as those of the original study, but had to be adapted to the locale of the replication studies. Whereas the first study replicated the original study's procedure, the next two studies extended it by employing increasingly more rigorous methods to examine the relationships between visual aesthetics and perceived usability.

Study 1 - Replication

Study 1 was an exact replication of K&K's method and stimuli with the exception that the Japanese stimuli had to be adapted to running the experiment in Israel. Most of the adaptation included the translation of the labels of certain controls of the ATM machine (e.g., the Confirm, Cancel, and Correction buttons). This part was quite simple, but there were two types of challenges. First, the original materials had to be reconstructed because of incompatible hardware and software. Second, while literal translation of the basic controls was straightforward, other parts of the interface were unique to Japan and were unfamiliar to Israeli users. For example, the original designs contained a large element depicting a feminine figure. This figure was

unique to Japanese ATMs. Israeli ATMs contained no similar element and it was feared that its inclusion would be met with skepticism (or worse). Thus, to prevent negative reactions on the one hand and to preserve the overall design layout on the other hand, the figure was replaced with a visual element of the same size, but which displayed an hour glass (see Figure 2, taken from Tractinsky, 1997).

Following the reconstruction of the stimuli the study followed the same procedure used in the original study.

Study 2 -Methodological Improvement I

Study 2 tested whether the results from the original study and its replication in Study 1 resulted from a method bias due to the fact that responses to the aesthetic and to the usability items were collected at the same time while the participants viewed the same design. That method carried the risk that the proximity of the measures would artificially inflate the correlation between them. To alleviate part of the concern, the study's procedure was modified. The 26 designs were displayed in two separated rounds. The order of presentation of the designs was randomized within each round. The order of evaluating beauty and ease of use was counterbalanced between two groups of participants.

Study 3 -Methodological Improvement II

In the original study and in the first two replication studies, the designs were presented to large groups of participants on a common screen, using a slide projector. In Study 3 the designs were presented on a computer screen by a program that also collected the

participants' responses. The use of computerized program allowed to further reduce potential biases by presenting the designs and the items measuring beauty and usability in a completely randomized order. The

differences between the three replicating studies are presented in Table 1.

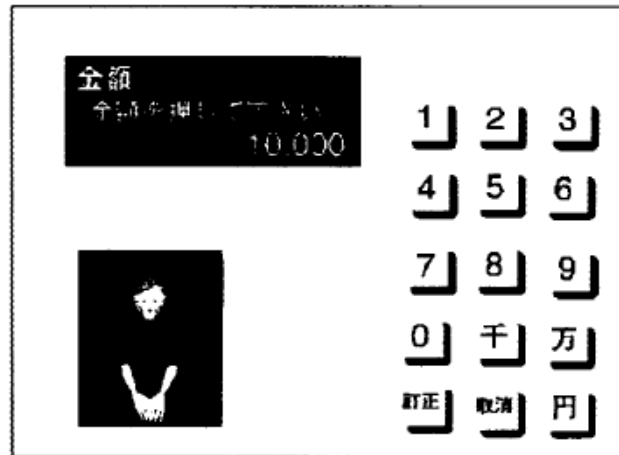


Figure 1(a). An original Japanese interface, rated high on apparent usability and aesthetics.



Figure 1(b). The equivalent Israeli interface, rated high on apparent usability and aesthetics.

Figure 2. Comparison of an original design and its counterpart in the replicating study

Study 1: Replication N = 104	Study 2: Improved method (1) N = 81	Study 3: Improved method (2) N = 108
Same procedure as original study. Designs adapted to fit local language/culture.	Items measuring beauty and usability were separated and their order of presentation was counter-balanced.	(a) Stimuli (designs) and measuring items were completely randomized. (b) Participants were seated individually in front of a computer.

Table 1. Summary of differences between studies of the replicating research.

Results

The three replicating studies yielded results similar to the original study in terms of the correlations between perceived (apparent) usability and six of the seven design guidelines, and most importantly, between perceived usability and evaluations of the designs' aesthetics. As can be seen in Figure 3, the basic findings remained unaffected by methodological improvements. If anything, the correlations between perceived usability and beauty were even higher in the replicating studies, demonstrating the robustness of the original findings.

Conclusion

The consistent results across cultures and following methodological improvements lent credibility to the findings of the original study. The original study and its replication opened up a new and lively research area in HCI regarding the role of visual aesthetics in HCI, and regarding its antecedents and consequences.

Variable	Correlations with Apparent Usability			
	KK	Exp. 1	Exp. 2	Exp. 3
AESTHETICS	.589	.921	.832	.920
DISTANCE	.000	.001	-.042	-.129
KEYPAD TYPE [#]	.730	.671	.751	.760
GROUPING	.075	-.462	-.529	-.667
SEQUENCE 1 [#]	.113	.352	.197	.397
HAND-DOMIN	-.127	-.002	-.125	-.203
SEQUENCE 2	-.306	.233	.137	.153
SAFETY	.137	-.019	-.006	-.061

Table 1. Correlations (bold: $p < .01$) and coefficients of contingency ([#]) of aesthetics and seven inherent usability variables with apparent usability for the experiment in Japan (KK) and for the three experiments in Israel.

Figure 3. Correlations between the design variables and perceived (apparent) usability in the original study and the three replication studies, as reported in Tractinsky (1997).

References

- [1] Kurosu, M., and Kashimura, K. (1995). Apparent Usability vs. Inherent Usability: experimental analysis on the determinants of the apparent usability. Conference Companion on Human factors in Computing Systems, Denver, USA, pp. 292-293.
- [2] Tractinsky, N., (1997). Aesthetics and apparent usability: empirically assessing cultural and methodological issues. ACM CHI Conference Proceedings on Human Factors in Computing Systems (CHI 97), pp. 115-122.

Replicating and Extending a Facebook Uses & Gratifications Study: Five Years Later

Tasos Spiliotopoulos

Madeira Interactive Technologies Institute
University of Madeira
Campus da Penteadá, Funchal, Portugal
tspiliot@gmail.com

Ian Oakley

Madeira Interactive Technologies Institute
University of Madeira
Campus da Penteadá, Funchal, Portugal
ian.r.oakley@gmail.com

Abstract

Social media change rapidly: new technological features become available and new communication practices emerge at a seemingly ever-accelerating pace. These dynamics raise questions about the validity of applying findings from past research to understand current systems. This paper explores this issue by a 2012 replication and extension of a prominent 2007 Uses and Gratifications (U&G) study on Facebook. The current study effectively built on the previous work by employing the same questionnaire items to measure

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.

and determine gratifications for using Facebook. Reassuringly, there was a high degree of similarity. However, an open-ended question that allowed participants to expand on the suggested set of gratifications yielded a large number of suggestions, indicating that a more comprehensive U&G study on Facebook may identify novel motivations for use, reflecting the increased scale, reach, and functionality of the site. The original study was also extended with the collection of empirical, numerical data derived from the Facebook API describing detailed Facebook usage and personal network structure. Motivations, challenges, successes and limitations of the replication and its extension are discussed.

Author Keywords

Replication, Uses and Gratifications; social network sites; social networks; Facebook; privacy; computer-mediated communication.

ACM Classification Keywords

H.5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous.

Introduction

Social Network Sites (SNSs) exhibit wide popularity, high diffusion and an increasing number of features. Specifically, Facebook, which currently holds a prime

Questions for eliciting items to be included in a U&G study.

- What is the first thing that comes to mind when you think about what you enjoy most when using Facebook?
- What other words describe what you enjoy about using Facebook?
- Using single, easy-to-understand terms, what do you use Facebook for?
- What uses of Facebook are most important to you?

position among SNSs, has a continuously evolving feature set and one billion monthly active users [4]. Given this diversity, an effective way of understanding Facebook is by exploring motives for using the service via theoretical frameworks such as Uses and Gratifications (U&G) [3, 7].

U&G is a theoretical framework for studying motives and outcomes – fundamentally, the “how” and “why” – of media use [3]. A typical U&G study employs a survey instrument (or occasionally interviews [8] or focus groups [2, 10]) for the collection of all relevant data. However, researchers have argued that more data-driven methods for the collection of U&G data can enhance the analytical power of the approach [7]. Apart from enabling the collection of a much larger set of data, the benefits of a data-centric study that follows a computational approach to measuring Facebook use would include freedom from issues such as recall bias [1], interviewer effects [6], and other sources of measurement error that may accompany survey research (see [5]), and assure the collection of accurate measures of users’ activity, broken down by specific Facebook features. In fact, as a theoretical framework, U&G does not mandate that any particular empirical methods be used and, therefore, the inclusion of computationally captured data in the U&G framework of analysis is a viable option [7].

In our forthcoming CHI 2013 paper [9] we have aimed at expanding the analytic framework of U&G theory to include *network* antecedents, as well as a more comprehensive and accurate measure of Facebook *usage*. In addition, we expanded the methodological scope of U&G by combining a typical survey tool with data captured using the Facebook API. In doing so, we

built upon the results of a highly-cited Facebook U&G study by Joinson [3] conducted in 2007. Our study was designed so that it is not “yet another U&G study”, but purposely and explicitly builds on the findings of Joinson to the extent that it can be considered a replication and extension of that work. This paper presents an experience report based on this replication and extension.

The Original Study

In July 2007, Facebook had recently moved outside the US-academic environment and had 30 million signed users. In his paper, Joinson employed a two-stage approach to studying uses and gratifications.

In the first stage, 137 Facebook users were asked to generate words or phrases to describe how they used Facebook, and what they enjoyed about their use. The questions used for this are shown in the sidebar.

These phrases were coded into 46 items, which were completed by 241 Facebook users in stage 2. In particular, participants were asked to rate, using a 7-point Likert scale, the 46 uses and gratifications derived from stage 1 using the metric, “How important are the following uses of Facebook to you personally?”. The scale was anchored at 1 (very unimportant) and 7 (very important). Participants also completed an item related to their use of Facebook privacy settings, specifically if they had changed the default settings, and if so, the degree to which they had made them more private or more open.

Factor analysis identified seven unique uses and gratifications: *social connection, shared identities, photographs, content, social investigation, social*

network surfing and status updating. Of the 46 items used in the factor analysis, 4 did not load on any of the factors and 14 did not meet factor purity criteria and were discarded, thus leaving a total of 28 items to describe the factors. User demographics, site visit patterns and the use of privacy settings were associated with different uses and gratifications.

Our Study and How it Compares

The goal of our study was two-fold. First, to combine the established framework of U&G theory with detailed usage and network data captured from an online social network service. Second, considering the dynamic and evolving nature of Facebook and the continuous introduction of new features, we aimed at investigating the extent to which the uses and gratifications identified in the 2007 study stand the test of time. For both goals, a direct comparison with the results of the previous study was deemed desirable and it was decided to build on those results instead of starting a U&G study from scratch. However, we were not explicitly interested in replicating the study as faithfully as possible (e.g., for validating the results), but simply using the same factors in our analysis because we considered that the two-stage process that was employed ensured accuracy and comprehensiveness of the identified items. Thus, we skipped the first stage of Joinson's study and instead utilized the 28 items he originally identified in a replication and extension of the second part.

In our study, participants were recruited with a request to complete an online survey. Recruiting was done differently than in Joinson's study, with approximately 1/3 of participants being recruited through posts on social network sites, 1/3 through posts to online

forums, mailing lists and online study repositories, and 1/3 through a Facebook ad campaign. Participants had to explicitly click a link to login with their Facebook credentials and access the survey, which is an equivalent action to installing a Facebook application. This combination of recruitment methods led to a sample that was more diverse in terms of demographic and geographic distribution, compared to Joinson's and to similar studies that typically take place within universities and study students. Since motives for Facebook use will likely vary substantially across cultures, ages, and educational backgrounds, the diversity of the sample used in this work may better match the traditionally exploratory nature of U&G studies. However, we should acknowledge a higher self-selection bias in our sample, since participants had to login with their Facebook credentials. On the other hand, this same process may have discouraged spurious participants (e.g., careless, dishonest, or mischievous web surfers). The size of our sample (208 participants) is comparable to Joinson's.

After logging in, participants were directed to an online survey capturing demographics and presenting 28 questions regarding their gratifications from Facebook, corresponding to the items identified by Joinson. Two questions examining attitudes towards privacy similar to Joinson's were also employed. Finally, participants were given the opportunity to expand on the suggested set of gratifications by answering an open-ended question that asked "Are there any other ways (not mentioned above) that you use Facebook for?".

In the meantime, the Facebook API was used to access a range of usage information for each participant. This included 11 variables, such as number of status

Suggestions for items to be included in future Facebook U&G studies.

Keeping up with news in general, keeping up with news from specific locations, keeping up with news from specific online news sources, following music bands, following specific news sources, following certain personalities (celebrities), following certain personalities (work-related), following organizations (e.g., theaters, clubs), entertainment and time-passing by following links suggested by friends, sending messages, remembering birthdays, promoting work, sharing/viewing videos, sharing music, chatting, video chatting, using email, maintaining professional relations, personal image control, organizing around school homework, seeing who is in a relationship with whom, linking to and promoting personal blogs, running Facebook Pages to connect with people with similar interests or fans.

updates made, likes given, check-ins made, and groups joined. In addition, the participant's Facebook friendship network was also collected enabling the calculation of 8 personal network metrics, such as size, density, and number of connected components.

An exploratory factor analysis was conducted on the 28 items, yielding seven factors, corresponding to motives for Facebook use, which are similar to those identified by Joinson. The differences between the factors identified in the two studies are in five items that did not load clearly, and the reinterpretation of the factor "Status updates" as "Newsfeed" to better reflect its constituent questions. In addition, a single item was moved to another factor.

Furthermore, the responses to the question "Are there any other ways (not mentioned above) that you use Facebook for?" yielded answers that suggest the inclusion of some new items to future studies, reflecting the dynamic and evolving nature of Facebook and the continuous introduction of new features. The most notable of these suggestions are shown in the sidebar. It is worth noting that some of these items were identified in the first stage of Joinson's study as well, but were discarded in the second stage due to not meeting factor purity criteria. Many others, however, are new reflecting new functionality in the service.

Extending the Study

The rest of our study followed a slightly different approach to the original. In Joinson's study, as happens in a typical U&G study, after the gratifications are gathered, the analysis examines the effect of the social/psychological antecedents and gratifications on the uses. However, since this analysis is purely

correlational, it is methodologically sound to reverse the directionality of analysis and attempt to predict the gratifications from the variables describing antecedents and uses, which is the approach adopted in our work. So, a series of multiple regressions were run with the seven motives (i.e., factor scores) of Facebook use as outcome variables, the Facebook usage metrics and network metrics as predictor variables, and the demographic variables as controls. Results showed that all three variable types in this expanded U&G frame of analysis (covering social antecedents, usage metrics, and personal network metrics) effectively predicted motives and highlighted interesting behaviors.

Two additional multiple regressions were run with the factor scores of the users as predictor variables and the answers to the two questions regarding privacy as outcomes. This aimed at further illustrating the power of this extended framework, by exploring the intricate nature of privacy in social media and drawing relationships between privacy attitudes (and acts) and measures of use and network structure.

Discussion on Replication

The results of U&G studies are typically reported in a way that facilitates replication; the data collection is clearly described and all the factors, items, and their loadings are reported. However, we are not aware of another U&G study that has been replicated (in social media, at least). In our case, there was no ambiguity about what happened in the first study and there was no need to contact the original author. Replicating the first stage of the original study might have produced some interesting results and possibly better highlighted the evolution in Facebook the past five years. However, doing this seemed out of the overall scope of our study

and could possibly lead to an uneven publication. The replication of the study was straightforward, but the extension required a bit more work, as more data were required. The differences in the sampling method and the data collection by the Facebook API had both advantages and disadvantages over the original study. Neither study can claim that its sample can adequately generalize to the Facebook population, but for different reasons each.

Conclusions

This paper presented an experience report based on this replication and extension of a U&G study on Facebook 5 years later. Our study effectively built on the results of a previous U&G study, by employing the items identified in the previous study to determine gratifications. The gratifications identified were very similar to those in the previous study, although it is not clear if it was expected since the same items were used for the exploratory factor analysis. However, an open-ended question that gave participants the opportunity to expand on the suggested set of gratifications yielded a large set of suggestions, hinting that a more comprehensive current U&G study on Facebook could identify new uses and gratifications, reflecting the evolution of the service the last few years. The original study was extended with the collection of a range of computationally collected data from the Facebook API covering Facebook usage and personal metrics, that effectively leverage prior research as a platform from which to expand the traditional U&G framework of analysis.

Acknowledgements

The work reported in this paper is supported by FCT research grant SFRH/BD/65908/2009.

References

- [1] Brewer, D. Forgetting in the recall-based elicitation of personal and social networks. *Social Networks* 22, (2000), 29–43.
- [2] Dunne, Á., Lawlor, M.-A., and Rowley, J. Young people's use of online social networking sites – a uses and gratifications perspective. *Journal of Research in Interactive Marketing* 4, 1 (2010), 46-58.
- [3] Joinson, A. "Looking at", "looking up" or "keeping up with" people?: Motives and use of Facebook. In *Proc. CHI 2008*, ACM (2008), 1027-1036.
- [4] Key Facts - Facebook Newsroom. <http://newsroom.fb.com/Key-Facts>. (Retrieved 13 January 2013).
- [5] Lewis, K., Kaufman, J., Gonzalez, M., Wimmer, A., and Christakis, N. Tastes, ties, and time: A new social network dataset using Facebook.com. *Social Networks* 30, 4 (2008), 330-342.
- [6] Paik, A. and Sanchagrin, K. Social Isolation in America: An Artifact (July 5, 2012). Available at SSRN: <http://ssrn.com/abstract=2101146>, (2012).
- [7] Papacharissi, Z. Uses and Gratifications. In M. Salwen and D. Stacks, eds., *An Integrated Approach to Communication Theory and Research*. Lawrence Erlbaum, 2008, 137-152.
- [8] Quan-Haase, A. and Young, a. L. Uses and Gratifications of Social Media: A Comparison of Facebook and Instant Messaging. *Bulletin of Science, Technology & Society* 30, 5 (2010), 350-361.
- [9] Spiliotopoulos, T., and Oakley, I. Understanding Motivations for Facebook Use: Usage Metrics, Network Structure, and Privacy. In *Proc. CHI 2013*, ACM (2013).
- [10] Urista, M.A., Dong, Q., and Day, K.D. Explaining Why Young Adults Use MySpace and Facebook Through Uses and Gratifications Theory. *Human Communication* 12, 2 (2009), 215–229.

NewsCube Replication: Experience Report

SidharthChhabra

School of Information
University of Michigan.
Ann Arbor, MI 48109 USA
sidc@umich.edu

Paul Resnick

School of Information
University of Michigan.
Ann Arbor, MI 48109 USA
prsenick@umich.edu

Abstract

We tested the robustness of a result demonstrated by Park et al, with the NewsCube system, that presenting suggested news articles related to a single story in clusters led to more exploration of articles and clusters [1]. We adjusted the apparatus to control for one potential confound in the original experiment, modified the experimental design to a within-subjects comparison to increase statistical power and allow assessment of subjective preference between the treatment and control interfaces, and switched from Korean to U.S. subjects to test generality. The results were only partially in agreement with the previous study. We reflect on the difficulty of drawing definitive conclusions when the original study and the replication differ in multiple ways. We also reflect on the challenges and value of conducting the replication as a learning exercise for a first-year doctoral student.

Author Keywords

Recommender; Diversity; Clustering; Replication; Experiment; Experience

ACM Classification Keywords

H.5.m. Information interfaces and presentation (e.g., HCI); Miscellaneous.

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

Introduction

We conducted a robustness test of a previous finding [1] for two reasons. First, we thought the previous finding had important design implications. Before it entered into designers' lore, we thought its reliability and generality should be checked. Second, one of the authors of this paper (Resnick) thought that a replication study would be a good learning exercise for the other author (Chhabra) as a first-year Ph.D. student. The idea was that it would be a chance to learn by example about experiment design, and about reporting findings in a scientific way, in the context of an interface that was known to be promising, rather than waiting to learn about good evaluation methods until after having devised an innovative interface of his own.

The Previous Study

Park et al conducted a lab experiment to find if presenting people meaningful clusters, at the sub-topic level, can lead to opinion (/political) diversity in what people read [1,4]. This was an important finding because diversifying exposure is good for society [5,6] but difficult to achieve through interface design [7,8]. Park et al's interface recommended articles in a sidebar while the subject was reading an article. Recommended articles were grouped together into clusters based on text similarity. If the subject had read from a cluster, then that cluster was grayed to mark that it had already been explored. They compared three different presentation methods: clustered, randomly clustered (i.e., articles randomly assigned to clusters) and unclustered. They found that people read more articles and explored more clusters in the clustered presentation than the unclustered presentation. On average, people read 4 articles from 3 clusters with the

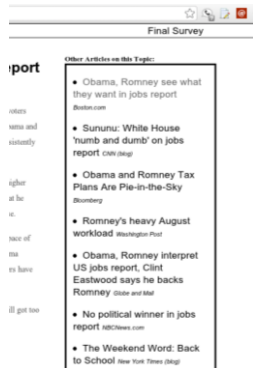
meaningfully clustered interface and 2.5 articles from 1.1 (unidentified) clusters with the unclustered interface. Random clustering did not produce any statistically significant effect compared to the unclustered presentation. They concluded that clustering was effective at encouraging people to explore multiple clusters, but only when the clusters were meaningful.

Our Study

To test the robustness of the finding that presenting sub-clusters of articles leads users to read more articles and clusters, we conducted a follow-on study. We re-implemented the NewsCube clustering algorithm [1] and the experimental apparatus. We then conducted a test with U.S. rather than Korean subjects. We used English language articles and allowed them to pick among many topics from the news of the day they came to the lab. Rather than directly replicating the original study apparatus and experimental design, we attempted to improve them slightly: we tried to identify and eliminate any potential confounds; and we switched from a within-subjects to a between-subjects design, to increase statistical power and to allow for elicitation of subjective preference for clustered vs. unclustered presentation.

Challenges

The first challenge was figuring out what the original study had done, to a sufficient level of detail to permit replication. Publications in the CHI conference are not expected to be replication ready: it is not among the review criteria and strict length limitations discourage inclusion of details that are important for replication but not for casual readers. For example, there were parameters in the clustering algorithm that were set



(a)



(b)

Figure 1 The sidebars for reading an article with unclustered treatment (a) and clustered treatment(b). The sidebar shows the articles suggestions with the read article(s) faded. In the clustered condition, explored clusters are grayed.

heuristically and neither the process nor the actual parameter values were documented in the original paper. It was not clear how the topics were chosen. It was not clear how many items were included in each cluster in the clustered presentation. Finally, the unclustered presentation was described as using the “Google News” interface. It was unclear whether this simply meant that it was unclustered within a news topic, as Google News was, or whether it literally used the same layout and interface elements used in Google News.

Fortunately, Souneil Park, the lead author of the previous study, was helpful; he answered many questions and discussed and critiqued our plans. We also had access to Park’s thesis (the second author of this paper was an external committee member for Park’s dissertation) which provided some details not found in the CHI publication. Without the additional documentation and help from Park, we would have been farther from a direct replication than we actually were, and in ways that neither we nor the scientific community would have been able to assess.

The second challenge was eliminating confounds. It turned out that Park’s unclustered treatment did literally use the Google News layout and interface, which differed in several ways from the custom interface used for the clustered treatment. First, the custom design was plain and had the feel of being done in the lab. More importantly, the unclustered treatment did not make recommendations in the sidebar while the clustered treatment did. In either condition, subjects could go back to a top-level page to select an additional article, but only in the clustered treatment could they do so without returning to the top-level page. We

eliminated this potential confound by adding a sidebar to the unclustered treatment. For the clustered condition, we picked the four most recent from articles from each cluster to display, or all of the articles for clusters with fewer than four. For the unclustered condition, we picked the same articles but did not group them into clusters, instead simply sorting them in reverse chronological order. Fig 1 shows both sidebars for a story.

Unfortunately, in retrospect we realized that we did not eliminate all the important differences between the treatments. First, after a user read any article from a cluster, the entire cluster was grayed and the actual article was faded out. For the unclustered treatment, only the actual article was faded out. The intention was to draw attention away from the already-explored clusters, but the impact may very well have been to draw extra attention to them. Second, because not all of the articles in the sidebar fit on the screen, users had to scroll to see them. Only a few clusters were “above the fold” in the clustered condition whereas in the unclustered condition articles from more clusters may have been visible without scrolling.

We corrected a second potential confound until just before our planned first lab session. In the top-level page for a story, for each cluster there was a short snippet for first article and for others only the title. In the unclustered treatment, a snippet was shown for only the very first article, rather than one for each cluster. We presented our design to the Michigan Interactive & Social Computing group (misc.si.umich.edu). Someone pointed out that a subject could read more snippets in the clustered treatment than the unclustered treatment, which might

influence how many articles they clicked on. To overcome this confound, we eliminated the top-level page altogether: the user selected a topic from the list of possible topics by selecting an article and was sent directly to that article, thereafter using only the sidebar to select additional articles.

The switch to a within-subjects design also created a new potential confound: order and contamination effects. For example a plausible order effect would be that users would explore more articles and aspects on the first topic that they read about than the second, simply because the first topic was more interesting. To control for that confound, we counter-balanced the order: subjects picked whatever topics they wanted to read, but some subjects got the clustered treatment for their first topic and some for the second. A contamination effect would occur if subjects' usage of say, the unclustered treatment, was different depending on whether they had already experienced the clustered treatment. This is always a risk in a within-subjects design, and we did not rule out this potential confound.

A third challenge was power calculation, estimating how many subjects were needed in order to have, say, a 90% chance of finding a statistically significant difference between the two conditions (at the .05 significance level), given an assumption about the size of the true difference in outcomes, the effect size. In principle, for a replication, the previously detected effect size and the observed variances in outcomes should have provided the needed inputs for a power calculation. In practice, however, since our design was within-subjects and the original was between-subjects, we had to resort to the same kind of guesswork that is

usually done in power estimations. We ended up recruiting 40 subjects, 20 in each experimental condition.

Results

Table 1 presents the key findings. Subjects read more articles and spent more time using clustered than the unclustered presentation. This confirmed the generality of one part of the previous finding, with testing across many more sets of stories, with articles in English rather than Korean. Subjects also preferred the clustered presentation, though not overwhelmingly so (26-14).

However, subjects did not explore significantly more aspects of a story in the clustered than the unclustered treatment. By aspects, we mean clusters produced by the NewsCube algorithm. Thus, our results were not consistent the most important finding of the original paper.

Publishing a Replication

We wrote a full paper about the replication, with more details about the apparatus, results, and limitations described above and submitted it to CHI 2013. It received several high-quality, thoughtful reviews, none of which recommended it for publication. Reviewers picked up on the potential remaining confounds that we reported (graying and scrolling). More generally, given that our results were not fully consonant with the original findings, and that we had changed many things, from language and subject pool to specifics of the interface, we could not make a firm conclusion about the correctness or generalizability of the main finding. Reviewers argued that in order to make a real contribution, further work is needed (they had different

	Clus-tered	Un-clus-tered	t-test
#Articles read	3.3	2.4	t(39)=3.0 p=0.003
#Clusters explored	1.7	1.5	t(39)=1.8p =0.24
Time spent	227 secs	183 secs	t(39)=2.1p =0.04.

Table 1. Reading results.

suggestions about what further work), and that we should really consider this a work in progress rather than a completed study. We found this argument persuasive, submitted a work-in-progress paper/poster for CHI 2013, and have plans for follow-up studies to yield a more conclusive result.

One review also argued, essentially, that neither the original study nor ours has yet demonstrated that the original finding was replication-worthy. We framed it as replication-worthy because of the prospect that using clustered presentation could nudge people towards exposure to diverse viewpoints, which is a valuable social goal. In fact, however, the clusters of articles represented different textual aspects of a news topic (i.e., clustering was based on text similarity), which might not necessarily represent different viewpoints. In future work, either an argument needs to be made that it's valuable to nudge people towards exploring multiple textual aspects of a story, or we will need to demonstrate that clustering on text similarity naturally leads to clustering on viewpoint similarity.

Discussion

Replication is an important ideal guiding the advance of science in any field. However, CHI papers are not yet ready for it. There is no expectation and no space in CHI publications for reporting sufficient detail to permit replication. Calculating heuristic parameters, and providing test databases and experiment observations are a few of such details. Perhaps, a norm and mechanism for published supplements providing fuller details would be good.

We also suspect that if our results had confirmed the original findings, a report of the study also would not

have been accepted, because it would not have been novel enough. Thus, for replication work to make a publication-worthy contribution in this field, it either needs to replicate and extend it, or it needs to show non-replicability and identify exactly why the original result did not replicate. This provides a limited incentive for any researcher to replicate and check a previous work. It's certainly not an easy path to a first publication for a student.

A replication study is still a valuable educational exercise for a first-year PhD student, and, if followed up, can yield a real contribution to accumulating generalizable knowledge. Through this replication, the first author gained a few lessons which he would not have learned otherwise. He gained an understanding of how to write research that can be replicated, providing every detail such that anyone can walk on the same path and conduct a similar experiment. It also taught him the importance of making available test cases and experiment data so that if in the future somebody wants to extend his work, he/she would be able to re-implement with confidence. Finally, we teach our PhD students about biases and threats to research validity. These concerns were driven home, however, to both student and advisor, when even after scrutinizing a previous study design for the better part of a year there were still potential confounds we identified after the fact in our own design.

References

[1] Park, S., et al., *NewsCube: delivering multiple aspects of news to mitigate media bias*, in *Proceedings of the 27th international conference on Human factors in computing systems 2009*, ACM: Boston, MA, USA. p. 443-452.

Teaching HCI Methods: Replicating a Study of Collaborative Search

Max L. Wilson
Mixed Reality Lab
University of Nottingham, UK
max.wilson@nottingham.ac.uk

Abstract

This paper describes the challenges experienced when replicating a user study that evaluated synergy in a collaborative search system. The original paper saw significant differences in collaborative performance, depending on the mode of collaboration. We were unable to replicate the findings, but experienced several challenges that created ambiguity and differences in the methods, which may have prevented us from doing so. These challenges and experiences, and their affect on our ability to replicate the findings, are described in detail.

Author Keywords

Collaborative search, Synergy, Replication

ACM Classification Keywords

H.5.3 [Group and Organization Interfaces]: Collaborative computing.; H.3.3 [Information Search and Retrieval]: Search Process.; H.3.7 [Digital Libraries]: User Issues.

Introduction

Hands on experience of replicating an experiment is often considered a good method of teaching [2]. For this reason, a cohort of 6 MSc students were asked to replicate a user study; to learn the methodological and analytical skills required to do so. Further, we hoped to confirm the findings for the benefit of the wider community. Based

Original Task Description

A leading newspaper has hired your team to create a comprehensive report on the causes, effects, and consequences of the recent gulf oil spill. As a part of your contract, you are required to collect all the relevant information from any available online sources that you can find.

To prepare this report, search and visit any website that you want and look for specific aspects as given in the guideline below. As you find useful information, highlight and save relevant snippets. Make sure you also rate a snippet to help you in ranking them based on their quality and usefulness. Later, you can use these snippets to compile your report, no longer than 200 lines, as instructed.

Your report on this topic should address the following issues: description of how the oil spill took place, reactions by BP as well as various government and other agencies, impact on economy and life (people and animals) in the gulf, attempts to fix the leaking well and to clean the waters, long-term implications and lessons learned.

upon the interests of the staff and students involved, we chose to replicate a user study of the synergetic effect experienced by users searching in collaboration, originally carried out by Shah and Gonzalez-Ibanez [5], herein referred to as the original researchers.

The original researchers studied their own collaborative search software (Coagmento¹), which had been evaluated previously [6], to examine synergy between collaborators in different group orientations. These orientations, as the primary independent variable, were co-located (same computer), co-located (different computers), and remotely located (different computers); individual searchers, automatically paired post hoc, were used as a baseline. The paper further contributed to the issue of evaluating synergy in collaborative search, by presenting new applicable measures. This focus on measures provided additional learning benefit to the MSc students involved.

The MSc students were given an entire semester to coordinate and run the study, and had each had to write about the results and the experience for their primary assessment. Support from the original researchers had been previously arranged by the staff.

Challenges Faced and Decisions Made

Significant challenges were faced throughout the replication attempt, from setting up the study, running the study, and analysing the results. These are described in turn below.

Setup Challenges

There were three major challenges in the setup phase: software procurement, data capture, and task design.

- **Software Procurement** - Initially it was considered that the procurement of software would be very easy, as Coagmento can be easily downloaded from the website. After installing the software, however, we noticed several differences in the user interface to the system described in the original paper [5]. The original researchers told us their study was based on an earlier version of the software. At first, we decided to accept the difference in functionality and to report it as a limitation later if needed. The original researchers, however, agreed to try and roll-back their functionality and provide us with a version that matched the evaluated version. This was very generous of the original researchers, and not always an option for those wishing to replicate studies.

- **Data Capture** - After investigating which data must be captured for the study, we discovered that the original researchers captured the data at the server level. Again, we were faced with two options: video record the desktop and manually log the necessary data afterwards, or request access to the data from the original researchers. The original researchers were again generous and agreed to provide us with the logs.

- **Task Design** - One significant challenge we faced was task design. The study was based upon an open-ended exploratory recall task, based upon american political parties. Our third decision was whether we should keep the american political task focus, or choose a more temporally (since the political topic had become old) and culturally relevant task for the British university. Several alternatives were proposed before making the decision, and in the end a temporally and culturally relevant task was chosen that focused on the 2012 Olympics (see original and revised task descriptions in the margins). This decision was made because task relevance and

¹<http://www.coagmento.org/>

Revised Task Description

A leading newspaper has hired your team to create a comprehensive report on the causes, effects and consequences of the Olympic Games. As a part of your contract, you are required to collect all the relevant information from any available online sources that you can find.

To prepare this report, search and visit any website that you want and look for specific aspects as given in the guideline below. As you find useful information, highlight and save relevant snippets. Make sure you also rate a snippet to help you in ranking them based on their quality and usefulness. Later, you can use these snippets to compile your report, no longer than 200 lines, as instructed.

Your report on this topic should address the following issues: Impact on economy of host countries (people and animals), long-term implications on the host country, conditions and voting policy to become hosting nation and the next host country and their preparations to host the games.

inherent motivation are considered key factors in creating good work tasks for user studies [7, 1].

Running the Study

There were three major challenges in the process of running the study: the experience of the research team, the financial support for incentives, and time limitations.

- Research Team - As this replication was being used to teach new MSc students about the process of running a study, the first and most obvious challenge is that the study is being run by inexperienced researchers. This challenge was further confounded by the necessity to teach many students at once. In this case, the original study was performed by one experienced phd student, but the replication was carried out by 6 novice MSc students. Each MSc student required experience at designing study materials (like questionnaires), handling participants, and analysing the results. This means that there was likely to be a high variance in each of the stages. To reduce variance, one final protocol was selected from each of protocols submitted by the students. However, there were not many constraints, apart from a default script, in terms of how, where, and when the researchers carried out the study with their participants.
- Financial Support for Incentives - As part of a taught module, rather than a funded research project, the students had to design alternative incentive methods. In the end, they choose a prize draw for a single prize (provided by the staff), but of a value much lower than a £10 voucher for each participant. There is some related work (e.g. [4]) into the style of different incentive structures, but the effect in this case was not clear.
- Time limitations - Also driven by the taught-module based constraints, the students had a limited amount of

time to perform the study. Consequently, the students had to make a decision, also relating to the financial limitations, about how many participants to include in the study. The students managed 40 participants in the timeframe, rather than the 70 involved in the original research.

Analysing the Results

There were two major challenges in the analysis phase: data processing and data analysis.

- Data Processing - The main challenge experienced in the analysis section was around the pre-processing of log data for analysis. The original researchers, for example, removed search engine result pages from their analysis of diverse website coverage, but the exact set of URLs considered as search engine results pages was implicit rather than explicit. In fact, any form of log processing and filtering in such a study would be a possible source of variance in user studies, unless the exact rules are accessible to the replicating team. One challenging example is whether to include both a user's typo and then their correction in analysing log data. In our own experiment, we created filters to achieve the same goals as reported in the paper, but we could not guarantee the exact same data would be filtered as the original research, given the same log; these elements of research methods are extremely difficult to comprehensively report in research publications.
- Data Analysis - With many methods, there are many variations on how to apply methods. In the case of this study, it was ambiguous as to how the data from the NASA Task Load Index (TLX) [3] was analysed. Many studies remove physical effort from the scale, as using a computer does not lend itself to variation in the physical

effort questions. In this case, it was unclear as to exactly how the NASA TLX was applied, including as to whether pair-wise comparisons were made.

Study Outcome and Discussion

The outcome of our replication attempt was that we could not replicate any of the original findings, as we hope may be reported in detail in a future publication. In summary, we saw no difference between the different measures, where the original researchers found a number of differences. However, there are many possible reasons for the differences, where we'll begin with the limitations of our replication attempt.

Limitations of our Replication

Although we were somewhat privileged to have the support of the original authors, we also had several limitations in our attempt:

- Researchers - our study was performed by 6 novice researchers, who each took part in running the study, with different individual abilities
- Participants - we had fewer participants (40 instead of 70), but from a similar academic population
- Participant Motivation - as part of a teaching module, participants were volunteers found by the MSc students, and were not motivated in the same way as original study
- Software - although the original researchers provided rolled-back software for the study, the process of rolling back introduced bugs that sometimes made the software unresponsive

Possible Causes of Different Findings

There are many reasons, including those listed above, that may have affected the outcome of our results, and

prevented us from getting the same findings. Reflectively, it's hard to estimate which element would have likely had the biggest impact on our attempt to replicate the study. First, the performance of the software, after being rolled back, was not ideal and this alone may have obstructed the synergetic effect seen by the original researchers. Second, the study was performed by several novice researchers, who may simply not have performed the study effectively. Third, the differences in the number of participants and the lack of voucher-based motivation could have limited the performance of participants. Fourth, task design has been seen to have a large affect on task outcome, and so perhaps your culturally and temporary relevant task may have not have been suitable. Finally, the processing of data for the analysis could have been simply different. Having some different or more comprehensive filtering rules may have led to significant differences in the measures.

Implications for RepliCHI

We chose to report this HCI replication, despite being focused on a user study not published at an HCI venue, because of the sheer number of issues that it highlighted for a community that wants to better support replication. Our specific example leaves many open questions that we may wish to investigate:

- What should we do when presented with different software versions from the original study?
- Should we use original tasks? Or is it acceptable to replace them for increased temporal/cultural relevance?
- Where data processing is involved, how should we best support others who wish to replicate our studies?
- If we want to recommend replication as a form of

teaching, what are the consequences of using groups of novice researchers?

- If we can't overcome these challenges, is there any value in replicating the studies?

Overall, the students experienced many challenges in trying to replicate the study, but learned a lot about study design and paper writing by doing so. For these educational reasons, the replication attempt provided a lot of value to the students. In terms of confirming the original study, we were unable to confirm the results, but were of course unable to disprove them also. This is perhaps a final challenge and discussion point for replication in HCI: we need to decide what we take away from studies that cannot replicate findings, and what value we have from understanding them. From this experience report, we hope that researchers may learn about several decisions that they may likely have to make when performing replications, and perhaps make more informed choices when the time comes.

Acknowledgements

We'd like to thank the original authors, Chirag Shah and Roberto Gonzalez-Ibanez for their support: providing software and advice for the replication.

References

- [1] Borlund, P. The concept of relevance in ir. *Journal of the American Society for information Science and*

Technology 54, 10 (2003), 913–925.

- [2] Frank, M. C., and Saxe, R. Teaching replication. *Perspectives on Psychological Science* 7, 6 (2012), 600–604.
- [3] Hart, S. Nasa-task load index (nasa-tlx); 20 years later. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 50, SAGE Publications (2006), 904–908.
- [4] Musthag, M., Raj, A., Ganesan, D., Kumar, S., and Shiffman, S. Exploring micro-incentive strategies for participant compensation in high-burden studies. In *Proceedings of the 13th international conference on Ubiquitous computing*, ACM (2011), 435–444.
- [5] Shah, C., and González-Ibáñez, R. Evaluating the synergic effect of collaboration in information seeking. In *SIGIR11: Proceedings of the 34th annual international ACM SIGIR conference on Research and development in information retrieval, July 24*, vol. 28 (2011), 24–28.
- [6] Shah, C., Marchionini, G., and Kelly, D. Learning design principles for a collaborative information seeking system. In *Proceedings of the 27th international conference extended abstracts on Human factors in computing systems*, ACM (2009), 3419–3424.
- [7] Wildemuth, B., and Freund, L. Search tasks and their role in studies of search behaviors. In *Third Annual Workshop on Human Computer Interaction and Information Retrieval, Washington DC* (2009).

Do lab effects transfer into the real-world? And should we care?

Petr Slovák
HCI Group
Vienna Uni of Technology
Austria
petr@igw.tuwien.ac.at

Paul Tennent
Mixed Reality Lab
University of Nottingham
United Kingdom
paul.tennent@nottingham.ac.uk

Geraldine Fitzpatrick
HCI Group
Vienna Uni of Technology
Austria
geraldine.fitzpatrick@igw.tuwien.ac.at

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.

Introduction

We report on two of our own studies, each of which has built on a laboratory based finding and explored if and how the effects played out in everyday settings. In each, we found effects that in some ways validated the prior lab studies, but each also pointed to very different implications for HCI than those which were suggested by the initial lab work. By lab-based work we mean here empirical studies that are tightly controlled and aim to uncover causal relationships.

We would like to use these examples to open a discussion within RepliCHI on whether or not the transferring of such lab findings into the field is a specific type of replication that is especially important for HCI research, as (i) we generally do want our systems/findings to transfer into field settings; and (ii) it is plausible to expect similar results to the ones in our studies when transferring other lab effects into the real-world. We expect that further discussion of this topic could well complement the existing “into the wild” literature in HCI that now focuses more on open-ended, in-situ exploration (e.g., see [1, 2]).

Moving effects from the lab to the wild

The two studies reported below come from different projects, but each builds on a finding that was previously rigorously analysed in the lab and seemed to be potentially useful in HCI.

Study 1: Feeling connected by sharing heartbeat

The first study picked up on a psychology experiment by Janssen et. al. [4]. Their work showed that sharing heartbeat between people increased feelings of intimacy and social connectedness. This effect was shown not only by statistically significant differences in questionnaire responses, but also by measuring changes in carefully chosen non-verbal aspects of the interaction. Implications of such findings in HCI could be, for example, the use of such an effect to design systems supporting mutual affection in couples living apart or helping to create stronger ties within families and other social groups etc.

To explore the potential for real-world application of the observed effect, we¹ developed a simple technology probe package based on a heart rate monitor belt paired with a standard laptop through a Bluetooth connection. Ten such packages were distributed amongst 5 couples, encouraging each couple to use the probe in any way they wish over the period of two weeks. Furthermore, we invited additional pairs of friends into our lab, let them experience the probe in social scenarios and interviewed them about their reactions and ideas. We analysed the interview and usage data by qualitative means, identifying two distinct effects appearing across most of our sample, and suggested interpretations as to why the effects happen. The results were then presented at CHI'12 [9].

What was interesting in the study with regards to this workshop is that while our results confirmed the initial study in many ways, the implications for HCI were strikingly different. For example, our participants reported feeling much “closer” to another when talking about situations closely resembling the lab study. However,

¹That is, Joris Janssen and two of the authors of this workshop paper (Geraldine Fitzpatrick and Petr Slovak)

people often felt actually “too close” in these moments, describing the feeling, e.g., “as if a stranger in an elevator was keeping eye-contact for a long time”. In other contexts, such as everyday use by the collocated couples, heart rate sharing did not have any effect at all (e.g. “as we are already close enough, this changes nothing”). Such results led us to suggest more specific contexts and situations where the effects of heart rate sharing could be used in positive ways, and better scope the potential applications of the original finding.

Study 2: Linking empathy to synchrony of bio-signals

In another study, we explored work done by Marci et. al. in psychotherapy and psychophysiology [6]. This work focuses on interaction between patient and therapist, and it links moments high in empathy to synchronised changes in skin conductance levels of the therapist/patient pair. For example, if the changes in skin conductance of patient and therapists were synchronised for a particular segment, external raters were more likely to rate such moments as high in empathy, it would also correspond to higher values in self-reported empathy etc.

Such a link could be of interest for HCI, e.g., as a novel indicator to embed in various affective computing systems, creating systems to support teaching of empathy for psychotherapy students, workplace etc. However, the original research was based on a very specific setting (therapy session) and participants with specific skills (therapist with many years of training to become highly empathetic). As such, we were interested to test how robustly the observed effects appear in the types of real-world settings that are of interest of HCI, but which are also often full of distractions and potential intervening variables that could not be controlled in real-world deployment.

We designed a study [10] in which pairs of friends discussed a topic of their choice in a public house during normal opening hours. The rationale was to test the robustness of the link in a setting that is more extreme (in terms of potential disruptions and intervening variables) than those needed for the potential applications. In other words, we argued that if the effect is robust enough to appear in a busy pub and for pairs of friends talking about any topics of importance, it is then more likely to appear also in a therapy students class, workplace setting or other potential application contexts.

Our results followed a similar pattern as in the first study: we have seen results that are in line with the original work, but the implications for HCI application of these have changed. For example, when we focused on interactions where participants were instructed to discuss their topics naturally, then thirty-seconds long video snapshots chosen purely on the basis of high synchrony showed also more empathy related non-verbal behaviour (as judged by independent raters). This fits with the prior lab results. However, we also found high synchrony in a condition where we asked one of the participants to ignore the other, i.e., where then little empathy could be expected. Such inconsistencies led us to suggest a re-interpretation of skin conductance synchrony – seeing it not only as an indicator of empathy, but potentially as a more general indicator of “mutual reactivity” (i.e., that people emotionally react to each other). Such reactivity then just happens to correspond well to empathy in the right contexts, such as therapy session or a discussion of two friends about an issue important for one of them. We were able to further support this hypothesis through other psychological literature such as [5].

Summary

To summarise, each of the two studies have shown that the expected effects can appear also in an uncontrolled, real-world setting, and are thus potentially robust enough for HCI applications. However, and maybe more importantly, each also clarified and better scoped the potential implications of the original finding for HCI.

Do we see a general pattern?

Stepping away from the two examples here, it does seem that, at least for results in psychology, transfer of effects from the lab to the field is far from an obvious claim. For example, Mitchell [7] shows in a recent meta-review that many lab effects either become much weaker when tested in the field, or even change direction entirely. Mitchell also shows how the extent of such “failures to transfer” differs among various sub-fields of psychology.

Can this be expected also of lab-based research in HCI? To our knowledge, there is little literature on this within HCI so far. It is also not discussed in the recent “into the wild” literature, e.g., [1, 3], which seems to have a more “open orientation towards finding out what happens and drawing design principles or recommendations about users’ reactions” [1].

We think it would be interesting to discuss in more depth how this focus on lab-to-field transfer of effects differs and complements the existing work on research “in the wild”. One immediate difference is the focus, i.e., whether a well-understood lab effect is robust enough to also appear in more realistic (and thus messy) conditions. Among other things, this will probably also raise methodological questions, as the main aim of such work is to test if an effect appears (thus pointing to more quantitative, experimental work), but in a setting where one cannot

control many of the potentially intervening variables. See Oulavirta [8] for an initial discussion of similar topic in the context of Pervasive computing.

Conclusion

We intended to demonstrate that examining whether the results of lab studies appear robustly 'in the wild' may be a specific kind of replication research, and one that could be of significant benefit to the CHI community. Drawing on our two studies, we saw that while the core effect did translate, the implications about how it might be used within HCI were changed markedly. We referenced additional literature in psychology suggesting that such results might also be expected for other lab-based findings.

Short Bio

Petr Slovak is a researcher and PhD student at the HCI Group at Vienna University of Technology. Drawing on his background in both psychology and computer science, his research focuses on support for teaching of empathy in medical and therapeutic settings, with specific interest in the use of biosensors.

Paul Tennent is a researcher at the Mixed Reality Lab at the University of Nottingham. He has worked on a number of systems designed to support the transformation of complex system log data into accountable, queryable objects that can be used in qualitative analysis. More recently he has turned to the analysis and representation of biodata with a particular focus on television and the public understanding of science.

Geraldine Fitzpatrick is Professor at Vienna University of Technology in Austria and heads the Institute of Design and Assessment of Technology. She is interested in how we design pervasive, tangible and ubiquitous technologies

to fit in with everyday contexts, with a particular interest in supporting social interaction and collaboration, and health and well being.

References

- [1] Brown, B., Reeves, S., and Sherwood, S. Into the wild: Challenges and Opportunities for Field Trial Methods. In *CHI '11*, ACM Press (May 2011), 1657.
- [2] Consolvo, S., Harrison, B., Smith, I., Chen, M. Y., Everitt, K., Froehlich, J., and Landay, J. A. Conducting In Situ Evaluations for and With Ubiquitous Computing Technologies. *International Journal of Human-Computer Interaction* 22, 1-2 (Apr. 2007), 103–118.
- [3] Hutchinson, H., Hansen, H., Roussel, N., Eiderbäck, B., Mackay, W., Westerlund, B., Bederson, B. B., Druin, A., Plaisant, C., Beaudouin-Lafon, M., Conversy, S., and Evans, H. Technology probes: inspiring design for and with families. In *CHI '03*, ACM Press (New York, New York, USA, 2003), 17—24.
- [4] Janssen, J. H., Bailenson, J. N., IJsselsteijn, W. A., and Westerink, J. H. Intimate Heartbeats: Opportunities for Affective Communication Technology. *IEEE Transactions on Affective Computing* 1, 2 (July 2010), 72–80.
- [5] Levenson, R. W., and Ruef, A. M. Physiological aspects of emotional knowledge and rapport. In *Empathic accuracy*. 1997.
- [6] Marci, C. D., and Orr, S. P. The effect of emotional distance on psychophysiological concordance and perceived empathy between patient and interviewer. *Applied psychophysiology and biofeedback* 31, 2 (June 2006), 115–28.
- [7] Mitchell, G. Revisiting Truth or Triviality: The External Validity of Research in the Psychological

Laboratory. *Perspectives on Psychological Science* 7, 2 (Mar. 2012), 109–117.

- [8] Oulasvirta, A. Rethinking Experimental Designs for Field Evaluations. *IEEE Pervasive Computing* 11, 4 (Oct. 2012), 60–67.
- [9] Slovák, P., Janssen, J., and Fitzpatrick, G. Understanding heart rate sharing: towards unpacking physiosocial space. In *CHI '12* (2012), 859–868.
- [10] Slovak, P., Tennent, P., and Fitzpatrick, G. Exploring physiological synchrony in everyday meaningful interactions. *Submitted for Interact'13* (2013).

Re-testing the Perception of Social Annotations in Web Search

Jennifer Fernquist

Google
1600 Amphitheatre Pkwy
Mountain View, CA 94043 USA
jenf@google.com

Ed H. Chi

Google
1600 Amphitheatre Pkwy
Mountain View, CA 94043 USA
edchi@google.com

Abstract

We evaluated the perception of social annotations designed via guidelines recommended by Muralidharan, Gyongyi, Chi, 2012. The initial study found participants noticed the annotation only 11% of the time with annotations shown below the search result snippet. Our refined study revealed that the proposed design with the annotation above the snippet increased noticeability to 60%. Replication studies are often iterative version of old studies, and this was no exception. The new study refined the protocol for measuring 'notice' events, and modified the tasks to include tasks that are more relevant to recent news articles.

Author Keywords

Annotation; social search; eyetracking; user study.

ACM Classification Keywords

H.5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous

Introduction

The abundance of information on the web suggests the importance of creating an environment in which users have the appropriate signals to make decisions about which search results are the most useful to them. As more of the web involves social interactions, they produce a wealth of signals for searching the most

interesting and relevant information. Much research has been done on modifying search ranking based on social signals for web pages [1][2][3][5][6], but how should we present the social signals for web search results? The most recent paper that we have found is the CHI2012 paper on social annotations by Muralidharan et al. [4].

Previous Research

Muralidharan et al. [4] studied the perception of social annotations appearing below search results, as in **Figure 1**. Consistent with prior papers, we use the term “social signals” to refer to any social information that is used to affect ranking, recommendation or presentation to the user. We use the term “social annotations” to refer to the presentation of social signals for an explanation as to why a search or recommendation result is presented. Thus, a social signal only becomes an annotation when it is presented to the user.

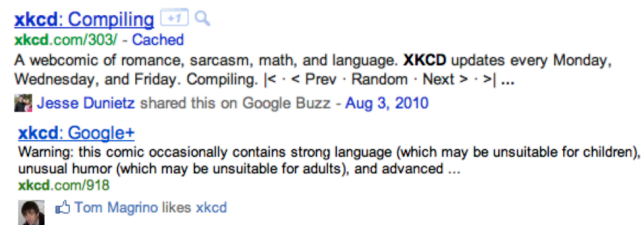


Figure 1: Example of older designs of social annotations. Image is from [4].

Study Protocol

Their first study had two parts: (1) In the first part, participants conducted 18-20 search tasks, randomly ordered. Half were designed so that one or two social annotations would appear in the top four or five results. The search results pages were presented as static mocks

that were generated before the study, customized for each participant.

(2) The second part consisted of a retrospective think-aloud (RTA) where they walked the participant through each task using the eyetrace data post hoc. During the interview, researchers checked noticeability by asking if the participants noticed the social annotations, either by them mentioning they saw them or being explicitly asked if they had seen them. During the RTA the researchers also obtained qualitative feedback about social annotations.

The second study compared the perception of multiple designs of social annotations. They varied profile image size (small, large), snippet length (1, 2, 4 lines), and annotation position (above, below snippet). For this study the same mocks were used for each participant, with customization only for customizing familiar names and faces of people in the annotations. In the second study, noticeability of the annotations was measured by counting the number of fixations.

Findings

In the first study, they found that only 5 of the 45 (11%) of the visible social annotations were noticed. In the second study, they found that there were fewer fixations on annotations when: the snippet length was longer; the image was smaller; and the annotation was below the snippet. They concluded that the optimal design for a social annotation is one with a large picture, above the snippet, with a short snippet length.

Our Method and Replication

We aimed to actually test the proposed annotation design guidelines from study 2 using live user data to

see if people notice the annotations more by using the method from study 1. Specifically, we wanted to test with live data that is relevant to participants (from their connections), as opposed to the static images used previously. An example of a social annotation with the new design is shown in **Figure 2**.

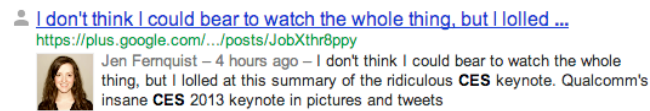


Figure 2: Example result with the new annotation design proposed by prior work. This annotation is above the snippet, has a large image and the snippet is less than 4 lines long.

Study Protocol

Experimental sessions consisted of 3 parts, the first two using essentially the same protocol as experiment 1's in the previous work, with some improvements.

PART 1: SEARCH TASKS

We designed planned 16-20 custom search tasks for each subject, at least eight of which were "social search" tasks designed to organically pull up social annotations. The 8 non-social search tasks were the same as used in the prior work.

In order to ensure that personal results appear for as many queries as required, we designed 2-4 additional social search tasks for each participant that were intended to bring up personal results. This way, if one social search task did not bring up personal results, we gave them the additional tasks to help ensure that they saw 8 tasks with personal results.

PART 2: RETROSPECTIVE THINK-ALOUD

After the search tasks, we immediately conducted a retrospective review of eye-tracking traces for search tasks in which subjects exhibited behaviors of some interest to the experimenter. Review of eye-tracking videos prompted think-aloud question answering about participants' process on the entire task, particular interesting pages, and particular interesting results.

Unlike Experiment 1 in Muralidharan et al. [4], we examined the eyetrace data directly by hand to determine noticeability, rather than through verbal feedback during the RTA tasks.

PART 3: THINK-ALOUD TASKS

Finally, participants performed two or three different search queries for which we determined ahead of time that should bring up relevant personal results. Here we gathered qualitative feedback on social annotations.

Results

In total, we collected eye-trace data for 153 tasks from nine subjects. Each eye-trace data for each task was analyzed by hand by an experimenter to understand: which positions contained personal search results; whether the search result was in the field of view in the browser; and importantly, whether the subject fixated on the result and/or the social annotation. This funnel analysis approach is different than the previous work's approach of asking participants if they noticed the annotations.

We discovered that participants fixated on annotations in 35 of the 58 tasks where they appeared (60%). This is a dramatic improvement over the 11% perception

rate of the Muralidharan et al. [4]. We account this difference primarily to the new annotation design.

Replication Discussion

Access to Previous Experimental Data. We were able to repeat the exact same tasks performed in the previous work but only because we share a co-author who had access to the data. If anyone else tried to replicate the study, they would not have been able to do so as effectively.

Temporal Challenges. Even though the search tasks were identical, because the study was conducted several months later, some of the task questions were no longer topically relevant. For example, one task asked “What is the website for the Google image labeling game?” At the time of our study, the website was no longer active. Similarly, the search task “Find some information about the Nevada law legalizing self-driving cars” brought up news articles from the previous summer, when Muralidharan et al. [4] conducted their research, since it was no longer recent news.

References

- [1] Bao, S., Xue, G., Wu, X., Yu, Y., Fei, B., and Su, Z. Optimizing web search using social annotations. In *Proc. WWW 2007*, 501–510.
- [2] Carmel, D., Zwerdling, N., Guy, I., Ofek-Koifman, S., Har’el, N., Ronen, I., Uziel, E., Yogev, S. and Chernov, S. Personalized social search based on the user’s social network. In *Proc. CIKM 2009*, 1227–1236.
- [3] Heymann, P., Koutrika, G., and Garcia-Molina, H. Can social bookmarking improve web search? In *Proc. WSDM 2008*, 195–206.

This raises a big issue for research replication: changing environments such as time or space. In our case, the tasks lost their relevancy over time. Researchers could help mitigate this by rewriting tasks so they are more relevant but still in the same vein as the original. For example, we could have written a different task that was more topical but would still be categorized as news. It must be decided which would cause the least amount of discrepancy for replication: maintaining the identical, less relevant task or rewriting a relevant task that differs from the original.

Iteration and Refinement. The primary difference in our protocol, measuring perception with fixation data rather than verbal confirmation, offered an improvement to the previous work.

Even with those challenges, we feel that we were successful in our replication efforts. We conducted an almost identical study to confirm the proposed improved design for social annotations and found a large increase in perception.

- [4] Muralidharan, A., Gyongyi, Z., and Chi, E. H. Social annotations in web search. In *Proc. CHI 2012*, 1085–1094.
- [5] Yanbe, Y., Jatowt, A., Nakamura, S., and Tanaka, K. Can social bookmarking enhance search in the web? In *Proc. JCDL 2007*, 107–116.
- [6] Zanardi, V., and Capra, L. Social ranking: uncovering relevant content using tag-based recommender systems. In *Proc. RecSys 2008*, 51–58.

Challenges of Replicating Empirical Studies with Children in HCI

Quincy Brown

Games+Mobile Play Learn Live Lab
Bowie State University
14000 Jericho Park Road
Computer Science Building
Bowie, MD 20715 USA
qbrown@bowiestate.edu

Lisa Anthony

UMBC
Information Systems
1000 Hilltop Circle
Baltimore, MD 21250 USA
lanthony@umbc.edu

Robin Brewer

UMBC
Information Systems
1000 Hilltop Circle
Baltimore, MD 21250 USA
brewer3@umbc.edu

Germaine Irwin

UMBC
Information Systems
1000 Hilltop Circle
Baltimore, MD 21250 USA
germaine.irwin@umbc.edu

Jaye Nias

Games+Mobile Play Learn Live Lab
Bowie State University
14000 Jericho Park Road
Computer Science Building
Bowie, MD 20715 USA
jayeacklark@aol.com

Berthel Tate

Games+Mobile Play Learn Live Lab
Bowie State University
14000 Jericho Park Road
Computer Science Building
Bowie, MD 20715 USA
TATEB0528@students.bowiestate.edu

Abstract

In this paper, we discuss the challenges of conducting a direct replication of a series of mobile device usability studies that were originally conducted with adults and older children (ages 7 to 17). The original studies were designed to investigate differences in how adults and children use mobile devices to touch targets and create surface gestures. In this paper, we report on a replication we conducted with young children (ages 5 to 7). We discuss several methodological changes that were needed to elicit the same quality of data from the replication with young children as had been obtained from the older children and adults. The insights we present are relevant to the extension of empirical studies in HCI in general to younger children.

Author Keywords

Child-computer interaction, touch interaction, gesture interaction, mobile devices, replication, empirical study.

ACM Classification Keywords

H.5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous.

General Terms

Design, Human Factors.

Introduction

In the context of research studies, children have often been viewed as vulnerable or ill-equipped and have been excluded from participation in studies due to concerns regarding informed consent, confidentiality, and the specialized attention or procedures required when conducting research with minors [1, 2]. To that end, previous research in human-computer interaction (HCI) has focused on including child participants by developing child-centered research methods and adapting protocols specifically for children [1, 3]. Though these efforts have increased the inclusion of children in HCI research, the use of separate protocols does not always allow for the direct comparison of findings between adults and children.

Here we present our insights from a direct replication of a series of studies of touch and gesture interaction on mobile devices that was first conducted with adults (18 years and older) and older children (ages 7 to 17) [6, 7, 8], and then replicated with younger children (ages 5 to 7). The goal of replicating these studies with younger children was to evaluate whether the same findings for older children and adults would hold for younger children. Though the studies were previously conducted with children as young as 7 years old, evidence from developmental psychology literature prompted us to include even younger participants: typically, as individuals mature from early childhood to adulthood, their cognitive and physical abilities also mature [4, 5]. Thus, the inclusion of younger children will allow for the comparison of patterns across all age groups, and support our overall goal of helping mobile application developers create more age-appropriate apps for children vs. adults, or even universally accessible apps.

Original Study Design

We have previously conducted three studies with adults (over 18 years old) and children (ages 7 to 17) [6, 7, 8] to investigate mobile device input and interaction differences between adults and children. The applications we used were designed specifically for these studies. Each participant completed a gesture task and target task. For the gesture tasks, participants used their finger to draw gestures (i.e., letters, numbers, symbols, and shapes) on the device screen. For the target tasks, participants touched square targets on the phone screen. A summary of the tasks from each study is given in Table 1. We also describe each task briefly to highlight the key points.

	Prelim. [6]	Study 1 [7, 8]	Study 2 [8]	Replica
No. Kids (Ages)	8 (7 to 11)	16 (7 to 16)	25 (10 to 17)	7 (5 to 7)
No. Adults (18+)	6	14	16	N/A
Target Task	Mini Target Task	Target Task	Target Task	Target Task
Gesture Task	No FB Gesture	FB Gesture	No FB & FB Gesture	No FB & FB Gesture

Table 1. Tasks and Studies.

Mini Target Task [6]

Square targets (43 in all) of four different sizes, large (26.4mm), medium (15.8mm), small (10.5mm) and very small (5.29mm), were displayed to the user one at a time. As the participant attempted to touch a target, the application logged the touch event. Participants were allowed one attempt per target only; touches were scored as *hits* or *misses*.

Target Task [7, 8]

The full target task used 104 targets of 4 different sizes: very small (3.175 mm), small (6.35 mm), medium (9.5 mm), and large (12.7 mm), in 13 different interface positions. This task incorporated *edge padding* for half the targets, which caused them to appear close to, but not on, the edge of the screen. The order of targets was designed to evenly represent all possible transitions between target positions and sizes, and no two consecutive targets had the same size or position. Unlike the mini target task, to advance to the next target, the participants had to successfully touch within the boundaries of the visible target. Therefore, multiple attempts for the same target were possible; touches were again scored as *hits* or *misses*.

Gesture Task – Feedback [7, 8]

Participants were shown a screen with text indicating which gesture to make and a “Done” button. Users used their finger to draw gestures on the device screen and press “Done” when finished. The complete gesture set (20 in all) included letters (A, E, K, Q, and X), numbers (2, 4, 5, 7, and 8), symbols (line, plus, arch, arrowhead, and checkmark), and geometric shapes (circle, square/rectangle, triangle, diamond, and heart). Participants were given a paper sheet showing what each gesture should look like, in case they were not familiar with every symbol by name (especially relevant for children). Participants entered an example of each gesture type one after another, and repeated this five times, yielding a total of six examples of each gesture type. As participants drew each gesture, a trace appeared under their finger of the gesture, but they were not able to edit their gestures.

Gesture Task – No Feedback [6, 8]

The no feedback gesture task was identical to the feedback task except participants did not see a trace of the symbol beneath their finger as they drew.

The Replica

We replicated Study 2 (conducted with older children and adults, see Table 1) using the same task applications: participants in the replicated study completed the Gesture Task – No Feedback, Gesture Task – Feedback, and the Target Task. So far, we have had 7 participants in this replication; three were 5 years old, one was 6 years old, and three were 7 years old. Of these participants, four were females, one participant was left-handed, and most self-rated their familiarity with touch input devices to be “average.”

Successes of Replica

The primary aspect of the protocol from the original study that was successful was the Target Task: in general, the 5 to 7 year olds were able to complete the Target Task without much difficulty. We believe this was because this task is very short and takes little time (about 1 to 2 minutes) compared to what is required to complete the six iterations of the gesture task (about 8 to 10 minutes). Furthermore, the Target Task required participants to perform an action (touching the interface) with which most children were familiar. In contrast, most of the children were not familiar with all of the gestures they had to draw in the Gesture Tasks and had to practice creating the gestures.

Limitations of the Replication

While the Target Task was a success, we encountered problems with the younger participants not completing all repetitions in the Gesture Tasks. Only 2 of 7 children

completed all iterations of the Feedback and No Feedback Gesture Tasks. The average number of rounds completed was less than 3 for the other children. With the majority completing so little of the task, we did not have enough data to be confident in results from a gesture recognizer (which needs enough data for both a training set and a testing set).

Comparison of Results from the Replica

Target Task – Misses

Table 2 shows the proportion of targets missed on the first attempt in the Target Task for all prior studies [6, 7, 8] and the replica with younger children. The 34% miss rate for the replica is higher than the Study 1 [7, 8] and Study 2 [8] miss rate, which we hypothesize is due to the younger age of the participants (the preliminary study [6] had a higher miss rate because the task only allowed one attempt per target).

	Adults	Children
Prelim. [6]	32%	46%
Study 1 [7, 8]	17%	23%
Study 2 [8]	16%	23%
Replica	N/A	34%

Table 2. Target Task miss results for all studies.

	Adults	Children
Study 1 [7, 8]	90% (FB only)	81% (FB only)
Study 2 [8]	91% (FB), 91% (no FB)	82% (FB), 85% (no FB)
Replica	N/A	46% (FB), 49% (no FB)

Table 3. Gesture Task recognition results for three studies.

Gesture Task

Table 3 includes the average per-user recognition results (computed for the replica using the open-source \$N\$ multistroke recognizer [9], as in prior work [7, 8]) for both Gesture Tasks across three of the studies. Both in spite of, and as a result of, the lower number of gesture samples collected so far during the replica, the replicated study recognition results are consistent with the overall trend we have found in our work that recognition rates are lower for younger participants. To ensure this finding is robust, we intend to explore ways to encourage children to complete the tasks so that we can examine this trend in more depth for the youngest children. We also hypothesize that the lower recognition rates may be attributed to the grade level of some of the participants (some 5 and 6 year olds had not completed first grade). Children who had been to school had more practice with handwriting and made gestures that appeared be more canonical (Figure 1).

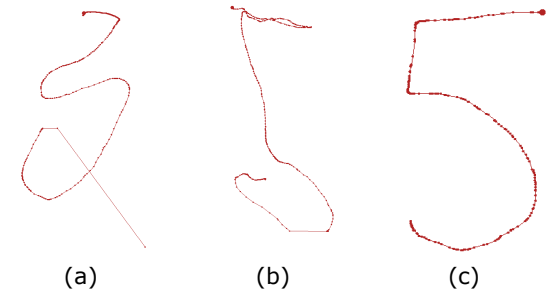


Figure 1. The gesture for the symbol '5' generated by a five (a), six (b), and seven (c) year old in the replica.

Reasons for Accounted Differences

In general, we found that the results of the replica were consistent with the original studies. However, we have identified four challenges areas with respect to younger

children not completing the gesture task portion of the study that could be useful for doing similar empirical replications with younger children in the future.

Motivation. All participants, adults and children, were compensated \$10 for their participation in Study 1, Study 2, and the replica [7, 8] (the preliminary study had no compensation [6]). Though financial compensation may motivate adults, we noted that the delayed financial compensation (receiving \$10 after the study vs. immediate rewards throughout the study) might not have been enough motivation for the young children in the replica.

Attention Span. We also noted that the young participants of the replica seemed less focused than the older participants from the original studies [6, 7, 8]. For example, they frequently told stories to the experimenter while completing the tasks, especially during the Gesture Tasks, and many asked for water or breaks during the session. Older participants in the original work did not exhibit this behavior [6, 7, 8].

Research Setting. All of the studies were completed in an academic usability lab with no windows [6, 7, 8]. This setting may not have been inviting and comfortable for the young participants of the replica. In the future, we plan to conduct studies in a more kid-friendly environment, such as a bright room with natural light and pleasant surroundings.

Acknowledgements

This work was partially supported by Department of Education HBGI Grant Award #P031B09020 and National Science Foundation Grant Awards #IIS-1218395 / IIS-1218664. Any opinions, findings, and conclusions or recommendations expressed in this

paper are those of the authors and do not necessarily reflect these agencies' views.

References

- [1] Punch, S. (2002). Research With Children: The Same or Different From Research With Adults? *Childhood* 9(3), 321-341.
- [2] Morrow, V. and Richards, M. (1996). The Ethics of Social Research with Children: An Overview. *Children & Society*, 10(2), 90-105.
- [3] Druin, A. (1999). Cooperative Inquiry: Developing New Technologies for Children with Children. *Proc. ACM CHI 1999*, 592-599.
- [4] Thomas, J.R. (1980). Acquisition of Motor Skills: Information Processing Differences Between Children and Adults. *Research Quarterly for Exercise and Sport* 51(1), 158-73.
- [5] Piaget, J. (1983). Piaget's Theory. In P. Mussen, ed., *Handbook of Child Psychology*. Wiley & Sons, New York, NY, USA.
- [6] Brown, Q. and Anthony, L. (2012). Toward Comparing the Touchscreen Interaction Patterns of Kids and Adults. *ACM CHI 2012 EIST workshop*, 4pp.
- [7] Anthony, L., Brown, Q., Nias, J., Tate, B., and Mohan, S. (2012). Interaction and Recognition Challenges in Interpreting Children's Touch and Gesture Input on Mobile Devices. *Proc. ACM ITS 2012*, 225-234.
- [8] Anthony, L., Brown, Q., Tate, B., Nias, J., Brewer, R., and Irwin, G. (In press). Designing Smarter Touch-Based Interfaces for Educational Contexts. *Journal of Personal and Ubiquitous Computing: Special Issue on Educational Interfaces, Software, and Technology*, to appear.
- [9] Anthony, L. and Wobbrock, J. O. (2012). \$N-Protractor: A Fast and Accurate Multistroke Recognizer. In *Proc. Graphics Interface 2012*, 117-120.

Replicating Residential Sustainability Study in Urban India

Mohit Jain

IBM Research Labs
Bangalore 560045 India
mojain13@in.ibm.com

Yedendra B. Shrinivasan

IBM Research Labs
Bangalore 560045 India
yshriv@in.ibm.com

Tawanna Dillahunt

School of Information
University of Michigan
4340 North Quad
105 S State Street
Ann Arbor, MI 48109
tdillahu@umich.edu

Abstract

Despite the global nature of problems such as rapid depletion of fossil fuels and water resources, most of the solutions being developed to address these issues are based on studies done in the developed world. We conducted a study of energy, water and fuel conservation practices in urban India, replicating the work of Dillahunt *et al.*, a qualitative study that explored the current practices, beliefs and attitudes of low-income households in two distinct U.S. locations. We used the same method, a photo-elicitation interview study, with 11 participants in Bangalore, India. Our study highlights *deep conservation* actions, which were influenced by the cultural context and different from the original work. Participants in our study shared motivations to conserve with participants in the previous study including scarcity, money, comfort and religion.

The purpose of this paper is to shed insight on our replication study. We discuss the purpose for conducting the replication study and describe the procedures we followed; we also provide information regarding access to procedures and data analysis techniques used from the original study. We discuss subtle differences in our procedure and how this may have affected our results and discuss key findings from our replication.

Author Keywords

Energy; Sustainability; Developing World.

ACM Classification Keywords

H.5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous.

General Terms

Human Factors; Design; Measurement.

Introduction

The goal of our study was to elicit a detailed picture of consumption and conservation practices and beliefs in Indian households. Like some of the prior work conducted in developed nations (primarily in the U.S. e.g. [1], [3], [7], [8]), we were interested in understanding motivations behind the conservation practices and challenges our participants faced around resource management. We decided to conduct our study in a developing nation as there was little information about whether or how prior results applied to other geographies, cultures, and socioeconomic groups. Further, we chose to focus on middle and high-income households because they consume resources in more diverse ways (e.g., own multiple types of appliances). Since our study was exploratory in nature, we chose to replicate a study conducted to understand energy consumption among low-income households in two U.S. locations [1].

Replication

The original study conducted used photo-elicitation interviews [1], which produces a different kind of information provoking feelings and memories. This information is not as easy to gather using standard interviewing techniques. Further, pictures provide a

focal point of conversation, which helps to alleviate any awkwardness an interviewee may feel [1]. Further, photo-elicitation interviews make it easy to agree on categories when analyzing data [2].

We analyzed the data using the same technique described in Dillahunt, *et al.* [1]. We coded and analyzed our interview data in an iterative fashion following methods taken from informed grounded theory [6].

Though photo-elicitation interview studies have been conducted in the past and well documented, in replicating the original study, we identified some aspects of the study that needed to be taken into account across various populations. For example, we made some changes in the protocol to factor in new contexts such as cultural differences.

Next, we discuss our method and differences that may have affected the results between the two studies.

Methodology

Prior conducting our study, we contacted the original researchers for their IRB material. This included recruiting detail, the surveys used to collect demographic information, and the specific script researchers read to participants. We made slight variations in the survey to accommodate for cultural context, such as the types of household appliances and transportation options. For example, we did not include dryers in our appliance list, as they were not as common among our population; we also added water heaters (Geyser) to the list. To understand conservation behavior we asked questions such as whether participants left the fan on to dry clothes, use

solar water heaters to heat water, conduct regular refrigerator maintenance, and/or use inverters (UPS). We also removed questions related to religion and spirituality as few participants were offended or felt uncomfortable answering those questions (though we made answering those questions optional). One such question was if they were motivated to conserve resources to protect God's creation. Access to this information helped in replicating the study method in its original form.

Differences in protocol

Despite being able to replicate all aspects of the study, there were some subtle differences that may have affected our results. These included the technology used to capture photos, payment, recruitment and the type of researchers conducting the study.

In the original study, participants used disposable cameras and at least one participant had never used a camera before the study. Our participants used either a digital camera or the cameras on their personal phones. Our participants had prior experience using the cameras. With these differences, participants using their own (digital) cameras may have felt more comfortable taking pictures and they may have been less concerned with running out of exposures. Though this unlikely had an impact on the results, it is a difference that should be considered.

The original study compensated participant for the time they spent during the interview. We had a different payment model. We did not pay our participants directly because we found during our interviews that participants were not interested in receiving payment. Instead, we paid our participants 2500INR to a charity

organization for every 50 participants to complete our online survey (the results of our survey were removed from our final paper submission).

The original study was conducted as a university study, whereas we were industry researchers conducting the same study. We were studying two distinctly separate populations, which makes it unclear how this may have influenced participant attitudes. As both studies were conducted in participant households, this may have alleviated any differences participants felt in terms of how comfortable they were in being interviewed. Our methods for recruiting were limited because we conducted our study as a private organization. As a result, we did not advertise publically—we relied on word of mouth and snowball sampling, which may have added bias to our participants.

From an internal organizational perspective, the "IRB" process for working with participants is slightly more difficult than in university settings. Industry is concerned about privacy issues such as IP; however, whether or not this is transparent to participants and affects their attitudes was not well understood.

Results

Many of our participants' conservation practices and motivations matched key categories of actions noted in the original study; however, as expected, the findings were not identical. We were able to contribute new categories and also leverage a vocabulary described in a more recent study, which provided evidence that the authors' framework generalized across different populations and cultures [3].

We also saw how our results generalized with the study we replicated and past studies of home energy consumption in developed regions. For example, participants in our study shared motivations to conserve with participants in past studies of typical [1], [3] and low-income households [1] including money, comfort and religion. Barriers to conservation such as money, comfort and safety also overlapped past studies. We highlighted two key differences between our findings and others in our final paper [5]. These include the impact of resource shortages (*scarcity*) and the value of *eco-feedback*.

When looking to generalize across lower-income U.S. households, our participants did not mention many common conservation behaviors. Our examples included re-using plastic drinking bottles for storing oils instead of buying dedicated containers, packing a family of 5 or 6 onto a single moped, and washing dishes using sand, ash, or coconut husk where water is in short supply—all findings unique to Indian culture. However, India has wide socio-economical, cultural, and demographic diversity, which makes it difficult to know exactly how broadly these findings generalize even within the country.

The major reason for differences among our work and the work replicated [1] is the shift in the cultural context. Hence we obtained many conservative actions, related to the Indian culture, but may not be relevant for developed countries.

Key Insights

We believe we can offer three key insights from our replication study. First, having access to scripts that describe the research method, the surveys conducted,

recruiting material, and access to a responsive original author, simplified our process. This information is often available in research Institutional Review Board documentation (IRBs); however, it is unclear whether this material is typically shared among researchers. Further, we are somewhat limited in our recruiting efforts due to the rigor required to advertise publically. This limited the types of participants that we could recruit and perhaps biased our results. Nevertheless, we found similarities between our results and the original study's results, as well as similarities between other home consumption studies.

Finally, in our study, we found the need to modify our demographic and baseline survey to account for cultural differences that existed between our study population, such as the types of resources used.

Discussion

Our replication was somewhat atypical as it was a replication of a qualitative study. However, our aim was not to replicate prior results. Our study was exploratory and we expected to see some conflicting results because of cultural and socioeconomic differences between the two populations; however, we anticipated some overlap as well. One topic for discussion is whether we can truly “replicate” a qualitative study. What exactly does it mean to replicate a qualitative study? Another question to consider is if using the same surveys was limiting in any way? We had to modify the survey based on cultural differences but was having the original material as a starting point a limitation?

- [1] Dillahunt, T., Mankoff, J., Paulos, E., and Fussell, S. It's not all about "Green": energy use in low-income communities. *Ubicomp 2009*, 255-264.
- [2] Harper, D. 2002. Talking about pictures: A case for photo elicitation. *Visual Studies*, 17(1), 13-26.
- [3] Pierce, J., Schiano, D.J., and Paulos, E. Home, habits, and energy: examining domestic interactions and energy consumption. *CHI 2010*, 1985-1994.
- [4] Rao, N., Sant, G., and Rajan, S.C. An overview of Indian Energy Trends. 2009. Prayas, Energy Group, Pune, India.
- [5] Shrinivasan, Y., Jain, M., Seetharam, D., Choudhary, A., Huange, E., Dillahunt, T., Mankoff, J. *CHI 2013*, (to appear).
- [6] Thornberg, R. Informed grounded theory. *Scandinavian Journal of Educational Research*, 56, 2012, 243-259.
- [7] Vyas, D. Domestic Artefacts: Sustainability in the context of Indian Middle Class. *ICIC 2012*, 119-128.
- [8] Woodruff, A., Hasbrouck, J., and Augustin, S. A bright green perspective on sustainable choices. *CHI 2008*, 313-322.
- [9] World Population Data Sheet 2012. http://www.prb.org/pdf12/2012-population-data-sheet_eng.pdf

Replicating and Applying a Neuro-Cognitive Experimental Technique in HCI Research

David Coyle

Interaction and Graphics Group,
Dept. of Computer Science,
University of Bristol,
Bristol BS8 1UB, UK
david.coyle@bristol.ac.uk

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.

Abstract

In cognitive neuroscience the sense of agency is defined as the as the experience of controlling one's own actions and, through this control, affecting the external world. At CHI 2012 I presented a paper entitled "I did that! Measuring Users' Experience of Agency in their own Actions" [1]. This extended abstract draws heavily on that paper, which described an implicit measure called *intentional binding*. This measure, developed by researchers in cognitive neuroscience, has been shown to provide a robust implicit measure for the sense of agency. My interest in intentional binding stemmed from prior HCI literature, (e.g. the work of Shneiderman) which emphasises the importance of the sense of control in human-computer interactions. The key question behind the CHI 2012 paper was: can we apply intention binding to provide an implicit measure for the experience of control in human-computer interactions? In investigating this question, replication was a key element of the experimental process.

Keywords

Replication; intentional binding; the experience of agency; evaluation methods

ACM Classification Keywords

H.5.2 [Information Interfaces and Presentation]: User Interfaces - Evaluation/methodology;

Intentional Binding

Repeated experiments have shown that voluntary human actions are associated with systematic changes in our perception of time [3]. The interval between a voluntary or intentional action and the outcome of such an action is typically perceived as shorter than the actual interval. For example, if a person voluntarily presses a button and this action causes an outcome - e.g. a beep - it is highly likely that the person will perceive their action as having happened later than they actually did (action binding). They are also likely to perceive the outcome as having happened earlier than it actually did (outcome binding). Patrick Haggard, the research who first identified this phenomenon, coined the term '*Intentional Binding*' to describe it, as it is contingent on several factors [3]. In the absence of outcomes people are found to more accurately report the timing of actions. For the temporal binding effect to occur, actions must be *intentional* and must lead to an *outcome*. Under these conditions our perception of the timings of actions and their outcomes become bound together temporally.

In the years since Haggard's first experiments, a large number of studies have validated and built on his initial observations. In and of itself this repeated experimentation highlights the importance of replication in cognitive neuroscience research. Based on this replication, a scientific consensus is now supported the conclusion that time perception in voluntary actions - and the binding effects associated with such actions - provides a robust implicit metric for the sense of agency. Higher intentional binding values correlate to a greater sense of personal agency.

Replication and application

Detailed descriptions of the experimental methods used to assess intentional binding are beyond the scope of this short paper. These details are available in the CHI 2012 paper [1]. Instead I will focus more broadly on the ways in which we replicated prior experiments and applied this metric.

Experiment 1

Neuro-cognitive experiments on intention binding typically focus on very simple interactions, e.g. a button press that causes a beep. My first experiment focused on the modality of the interaction. It asked if changes in the modality of an interaction lead to changes in the sense of agency. The experimental design closely mirrored procedures originally outlined by Haggard. One independent variable was manipulated: the input modality. We compared a traditional input device - a keypad - with a skin-based input device. The keypad replicated the input typically used in neuroscience research. In condition one the participant pressed a button on a keypad to cause a beep. In condition two - the skin-based condition - the participant caused a beep by tapping on their arm. The skin-based capture device was attached to the participant's left arm and they tapped this arm with their right hand. In all cases there was a fixed interval of 250ms between the participant's action and the beep.

Results showed that users experienced significantly higher intentional binding for skin-based interactions. Across 19 participants a mean binding of 42.92ms was observed in the button press condition. I.e. an interval of 250ms was perceived as 207.08ms. Importantly, this binding value is consistent with the results of prior

binding experiments that have used button inputs. In the skin-based condition participants experienced a total binding effect of 109.47ms. Here 250ms was perceived as 140.53ms. Given the correlation between intentional binding and the sense of agency, this experiment suggests that people experience a significantly greater sense of agency, or control, when they interact with technology via skin-based input, as compared with traditional keypad input.

More broadly speaking, this experiment provided empirical evidence that different interaction modalities can provide different experiences of control and ownership. In undertaking this experiment I believe it was essential that one of our input conditions - the keypad - replicated prior cognitive neuroscience research. This replication demonstrates that our experiment was administered effectively and lends strength and credibility to our findings. It also allows our results to be judged against and incorporated into the prior body neuro-cognitive research on intentional binding and the sense of agency.

Ultimately I hope the method we introduced can be used to investigate the sense of agency across a wide range of input modalities. For other researchers using this technique, I strongly recommend that replication (plus extension) of prior results again be a key element in the design of new experiments.

Experiment 2

Cognitive neuroscience experiments on intentional binding have typically examined voluntary and involuntary actions. From an HCI perspective, this might be considered an unnecessarily black or white disjunction. Many user interactions with technology are

more intermediate. In particular 'intelligent' user interfaces often seek to interpret and act on the intentions of the user. Here users' actions are voluntary, but the outcomes may be assisted. The second experiment in the CHI 2012 paper was designed to investigate users' sense of agency in interactions where a computer interprets their intention and helps them to achieve a goal. In this sense the second experiment diverged further for the interactions examined in prior cognitive neuroscience research. However, as in first experiment, we apply an experimental procedure that closely matched prior literature.

The experiment investigated agency in a machine-assisted point-and-click task. Using a mouse, participants were required to hit targets on a computer screen, as quickly and as accurately as possible. The computer provided assistance through an algorithm that effectively added gravity to targets, thereby making it easier for participants to complete the task. Hitting a target caused a beep. In each trial there was a random interval between hitting a target and the beep, and participants were asked to estimate this interval.

In the experiment we investigated four different assistance levels, which varied from no assistance to a very high, and very obvious, level of computer assistance. Results suggested that, up to a certain point, the computer could assist users whilst also allowing them to retain a sense of agency for their actions. However, we found that beyond a certain level of assistance users experienced a detectable loss in their sense of agency. This loss in agency occurred in spite of the fact that the computer correctly interpreted

users' intentions and assisted them in achieving their goal.

Our results suggest that for the assisted input algorithm we investigated - and possibly for assisted input systems more generally - there may exist a tipping point or sweet spot. This is the point at which a computer can help people and potentially maximise task performance - e.g. speed or accuracy - without significant detriment to the experience of agency. I find this possibility very intriguing. However I also believe further investigation, and further replication, is required to assess the generalizability of our initial finding. I am currently undertaking such research.

Conclusions

Alongside the issues discussed above, I have one minor comment on the CHI submission process. When I submitted the original CHI 2012 paper, I was very keen

to also submit the dataset for my studies. Under the 2012 submission system this was not possible. I understand that this issue was addressed for CHI 2013 submissions. This was a real step forward.

Citations

- [1] Coyle, D., Moore, J., Kristensson, P.O., Fletcher, P.C., & Blackwell, A.F. (2012) I did that! Measuring Users' Experience of Agency in their own Actions. ACM CHI 2012, 2025-2034.
- [2] Ebert, J.P. & Wegner, D.M., *Time Warp: authorship shapes the perceived timing of actions and events*. Consciousness and Cognition, 2010. 19 481-89.
- [3] Haggard, P., Clark, S., & Kalogeras, J., *Voluntary action and conscious awareness*. Nature Neuroscience 2002. 5(4).

Replicating Two TelePresence Camera Depth-of-Field Settings in One User Experience Study

Jennifer Lee Carlson

Sr. User Experience Researcher
Cisco Systems, Inc.
170 W Tasman Drive
San Jose, CA 95134 USA
jennicar@cisco.com

Mike Paget

Sr. Technical Marketing Manager
Cisco Systems, Inc.
170 W Tasman Drive
San Jose, CA 95134 USA
mpaget@cisco.com

Tim McCollum

Sr. Design Manager
Cisco Systems, Inc.
170 W Tasman Drive
San Jose, CA 95134 USA
tmccollu@cisco.com

Abstract

This paper describes an experience study to understand the user perceptions on two camera focus settings in a TelePresence room: limited- and infinite-Depth-of-Field. The results influence future TelePresence experience design.

Author Keywords

User experience; comparative study; TelePresence; Depth-of-Field; video conferencing codec; macroblocks; network bandwidth

ACM Classification Keywords

H.4.3. Information Systems: INFORMATION SYSTEMS APPLICATIONS: Communications Applications

General Terms

Human Factors, Experimentation, Design

Introduction

Depth-of-Field is a description of the focal characteristics within a captured image. It describes the sharpness of the image from the foremost to farthest areas on the z-axis within the cameras field of view. The cameras Depth-of-Field is determined by four key factors which were related to works in this study:

1. The proximity of the two lenses to the camera sensor and the cameras overall proximity to the subject, otherwise known as focal length.
2. The amount of light that is allowed to reach the sensor controlled by the aperture setting.
3. The duration at which light is allowed to pass through the aperture, which is called shutter speed.
4. Camera gain setting, which can increase perceived brightness in the image.

Two very common approaches to image capture produce very different resulting images under the same environmental conditions. The approach of limited Depth-of-Field is to limit the amount of focal area within the image to achieve controlled focal points. Generally this is an artistic decision for a particular aesthetic style. This has also been used in video applications such as cinema for the same artistic purpose. However, in video applications such as conferencing, the same image characteristics have been used for a completely different purpose [1]. Current video codecs used in conferencing systems apply an algorithm that defines what information is sent based on changed events rather than sending the entire image. The algorithm groups areas of information together in macroblocks and sends these chunked updates when an area of the macroblock has changed. This approach requires tedious preparation of the environmental conditions and the camera settings. In the case of the Cisco TelePresence System 3000-series (CTS-3xxx) system designs, they were purposely built for a dedicated room that was optimized for very high quality at a low network bandwidth. Therefore they followed the model of camera settings that provided limited Depth-of-Field.

On the other hand, the approach of infinite Depth-of-Field, where a controlled focal point is not established, is also common in both still image and video image capture. This approach captures more detail within the resulting image and requires the viewer (or end user) to determine their own focal points as they view and process the image. In a conferencing system this approach will capture objects within the camera's field of view, as they exist without the need to adjust the amount of sharpness. Such a system could require

additional processing power and bandwidth requirements but it doesn't require the same tedious attention to detail of the environmental conditions or the camera settings. Therefore this approach offers a more flexible deployment model for a wider range of conditions. In the case of the systems that utilized the Precision HD camera, such as the Cisco TelePresence 3-series (T3) system, they shared the same camera for both dedicated and multipurpose room systems. Therefore use of an infinite Depth-of-Field configuration was preferable to allow greatest amount of flexibility.

The two depth-of-field applications were largely based on technical and business reasons. What are the user experience impacts, if any, from the two camera settings in a TelePresence room? Our usability study was to answer the following questions:

1. Are users aware of the difference in the two camera focus approaches? If so, how do they differ?
2. Which approach feels more life-like to users? What made it more life-like?
3. Which approach do users prefer and why? Are there other considerations besides being life-like?

Methodology

In August 2011 the Cisco TelePresence User Experience team conducted a formal usability study in an immersive TelePresence room (see Figure 1).

The study replicated the two camera settings in the same TelePresence room to evaluate the user experience in the context of a meeting. During the session, users focused on the moderator, no documents were shared, and the room had sufficient depth and background to identify the moderator's unique location.

We conducted a total of 27 within-subject comparative usability study [2] sessions, with each session lasting approximately 15 minutes. All participants have experience with TelePresence.



Figure 1. Cisco TelePresence CTS-3000 System

Participants entered the TelePresence room containing three side-by-side HD screens. The middle screen was turned off during the entire study. The participant was seated in the middle of the room so that the left and right screens were the same distance from their seat. The left screen displayed an infinite Depth-of-Field, where both moderator and background were in focus. The right screen displayed a limited Depth-of-Field, where the moderator was in focus but the background was blurred at a noticeable level. After the second day of sessions (completed 15 participants) the background objects were switched completely to counter-balance any effects due to the background objects.

In this study, in order to evaluate the user experience impact from the two depth-of-field settings, it was

critical to make the other aspects of the images as similar as possible, such as the field of view and the subject matter within the frame. The two cameras (used for the CTS-3xxx and T3 systems) for which we wanted to test had very different physical characteristics. But it was important that users couldn't tell the different cameras by their physical appearances from the room. It was not possible to house both types of cameras within the same system. Therefore, one camera type was selected that fit the appropriate physical characteristics as well as possessed settings that could achieve both a limited depth-of-field and infinite depth-of-field.

Replicating the Depth-of-Field technique is relatively easy in some areas and difficult in others. The lighting and camera settings (hard and soft) can be replicated easily with this controlled environment. The actual focal settings are more challenging because we didn't actually measure the depth of field with any equipment. It was assumed based on camera and light settings, and by looking at the two set-ups subjectively. However, if we were to focus more effort on measuring the depth of field as to define the distance and amount of sharpness or blurriness, it could be more easily replicated. The other area that was challenging to replicate are the objects in the background. We setup similar background based on props we had available. We could define more parameters on those props to better replicate the testing.

Procedures

Participants were told they would have a conversation with a moderator via TelePresence to discuss their experience with TelePresence, provide feedback and rate their experience. Participants were not informed of

the difference in camera approaches until after they had separately provided feedback and rated both views. They looked at one view at a time until the very end of the session when they compared the views side by side.

The study itself was comprised of three separate elements:

Camera Setting 1

Participants were first presented with a view (segment) of the moderator on either the right or left screen (the order was reversed for every other participant to avoid potential order effects). After answering TelePresence-related questions for several minutes, participants were asked to rate the TelePresence session in terms of video quality and how lifelike it appeared.

Camera Setting 2

Then the view was switched to the opposite side of the room and the moderator moved to the displayed view to interact with the participant. After several minutes of additional conversation, the participants were again asked to rate the video quality and lifelike appearance of the view.

Comparisons

Participants were asked if they could tell any differences between the two views they just looked at. If there were any differences, how the two views appeared differently. Then they were shown both views - one at a time - and asked if they noticed any difference, or if they have noticed any other differences. At the end, participants were shown both views simultaneously so that they could make direct comparisons. Participants were asked to describe any differences they observed. If the participant could not discern a difference in background clarity, the

moderator explained the differences between the infinite and limited Depth-of-Field camera approaches. With this knowledge, participants then rated how appealing each view was, which view they preferred and why.

A 7-point scale rating scale was used for all rating questions, where 1 represented the 'worst' rating and 7 represented the 'best' rating.

Findings

The study has identified the following key findings based on participant behavior, feedback and preference ratings:

1. Approximately 93% (25 of 27) participants were unable to distinguish the camera focus approaches on their own without viewing the images side by side. Even after viewing the images side by side, only 37% (10 of 27) of participants were able to discern the difference in background clarity between the two views.
2. Between the two camera focus approaches, on average there were very minimal perceived differences in terms of being lifelike (5.93 for infinite Depth-of-Field vs. 5.86 for limited Depth-of-Field) and video quality (6.32 for infinite Depth-of-Field vs. 6.29 for limited Depth-of-Field.)
3. After understanding the camera focus difference: More participants (11 of 27 or 40%) preferred the infinite Depth-of-Field approach. Fewer participants (8 of 27 or 30%) preferred the limited Depth-of-Field approach. Almost one-third (8 of 27 or 30%) participants did not have a preference between the two approaches. On average the infinite Depth-of-Field view was rated slightly more appealing (5.93 for infinite Depth-of-Field vs. 5.52 for limited Depth-of-Field).

Conclusion

Camera's Depth-of-Field setting is not a significant experience differentiator for an immersive TelePresence room. Infinite Depth-of-Field could potentially provide a more lifelike experience and perceived as better quality.

Potential Future Work

This study was meant to be the first of a series of studies. We want to find out what degree of camera focus difference will be perceivable by most users. We also want to study and analyze how user's preferences for camera focus relate to the different types of meetings: such as an interactive brainstorming session, a round-table team meeting, a single-speaker presentation, or other types of meetings.

One hypothesis was that users who are more technical or goal oriented might show a stronger preference for limited Depth-of-Field because they might focus more on the people than their environment; users who are more artistic or context sensitive might show a stronger preference for infinite Depth-of-Field because they care more about the surroundings of whom they meet with. There wasn't any analysis on how the Depth-of-Field preferences relate to participants' job roles or personalities.

Acknowledgements

The authors thank all participants in Cisco Systems who participated in the study described here. We also thank Laura Borns of Cambridge Consultants for her note-taking and analysis assistance for the study; Kevin Nguyen and Rick AtKisson of Cisco Systems for their support in TelePresence room set-up for the study, and Chris Dunn of Cisco Systems, who initiated this research study and reviewed this submission.

References

- [1] O'hara, K., Kjeldskov, J., Paay, J., Blended interaction spaces for distributed team collaboration. In ACM Transactions on Computer-Human Interaction (TOCHI) TOCHI Homepage archive, Volume 18 Issue 1, April 2011, Article No. 3.
- [2] Sauro, J., Lewis, J.R. Quantifying the User Experience: Practical Statistics for User Research (2012, ISBN-10: 0123849683 | ISBN-13: 978-0123849687), 10-11.