Is Once Enough? On the Extent and Content of Replications in Human-Computer Interaction

Kasper Hornbæk[1], Søren S. Sander[1], Javier Bargas-Avila[2], Jakob Grue Simonsen[1]
[1]Computer Science, University of Copenhagen,
Njalsgade 128, DK-2300 Copenhagen, Denmark
[2]Google/YouTube, User Experience Research
CH-8002 Zürich, Switzerland
kash@diku.dk, s.s.sander@mail.dk, javier.bargas@me.com, simonsen@diku.dk

ABSTRACT

A replication is an attempt to confirm an earlier study's findings. It is often claimed that research in Human-Computer Interaction (HCI) contains too few replications. To investigate this claim we examined four publication outlets (891 papers) and found 3% attempting replication of an earlier result. The replications typically confirmed earlier findings, but treated replication as a confirm/not-confirm decision, rarely analyzing effect sizes or comparing in depth to the replicated paper. When asked, most authors agreed that their studies were replications, but rarely planned them as such. Many non-replication studies could have corroborated earlier work if they had analyzed data differently or used minimal effort to collect extra data. We discuss what these results mean to HCI, including how reporting of studies could be improved and how conferences/journals may change author instructions to get more replications.

Author Keywords
Replications.

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

INTRODUCTION

A replication attempts to confirm an earlier study's findings, sometimes in conditions almost identical to the original study, sometimes in conditions that differ with respect to manipulations, setting, or measures. Replications thereby help establish whether a finding can be repeated – increasing confidence in its validity – and describe the conditions under which it holds or fails. They also help weed out studies whose results cannot be validated by independent research teams or under slightly different conditions. Therefore, replications help generalize findings and ensure intersubjectivity [24], and are a key to making science cumulative. A recent special section in the journal *Science* called replications the "scientific gold standard" [21]; Bazerman [2] wrote that "in replication the private chimera becomes the communal fact".

In Human-Computer Interaction (HCI), the discussion of the extent and value of replications recurs. For example, a series of events at the ACM CHI conference ([51], www.replichi.org) has promoted replications. Another example is a much debated paper on usability evaluation [11] that discussed the lack of replication in HCI. It argued that reviewers do not value replications and that replications are often not published, despite their potential value in nuancing earlier results. Zhai et al. [53] reported a replication of a study of target expansion by changing how the original study presented target expansion and how errors in performance were analyzed. They motivated the need for this replication by writing "although rarely done in the field of human computer interaction, important research results should be replicated for soundness and sustained development of a research topic" (p. 179).

In fields outside HCI, studies have shown that the literature contains few replications [e.g., 8,26,46] though some argue that a substantial number of replications exist [e.g., 23,39]. The success of meta analyses in some fields, most prominently medicine, presupposes studies that replicate each other [19]. In some fields it has been shown that replications often fail to find support for the claim they were supposed to replicate [2,48], though also this finding varies across fields [cf. 23]. In several highly publicized cases, initial claims of spectacular results have not been reproduced in spite of determined efforts by several independent research teams. Examples include cold fusion [18] and the ability of certain microbes to sustain growth by substituting arsenic for phosphorous [43].

To our knowledge no study has systematically investigated replications in HCI. Thus we do not know (a) the extent of replications in HCI or (b) the content of those replications. The present paper reports a study of replications identified through systematically examining conference proceedings and journals. We code the extent and content of replications, analyze the replications found, and discuss the relation between them and the studies being replicated. We believe that our work will (i) strengthen discussions of replications in HCI by providing data on the extent and

CHI 2014, April 26 - May 01 2014, Toronto, ON, Canada
Copyright 2014 ACM 978-1-4503-2473-1/14/04…$15.00.
http://dx.doi.org/10.1145/2556288.2557004

content of replications, and (ii) assist in developing HCI-specific guidelines for replications that will facilitate more and better replications.

RELATED WORK

Replications have been much discussed in the literature on scientific methodology [e.g., 24,45]. Usually, replications are taken to mean the confirmation of the results of one study in a second, independent study [45]. Despite the importance of original, first discoveries in science, many have argued that there is too much emphasis on such discoveries [e.g., 20], and consequently less on replications.

The question of what constitutes a replication is hard to answer. In a strict sense, a study cannot be replicated because the investigator, setting, time of study, and instrumentation will differ [45]. In a less strict sense, most scholars rely on some classification of replications. Makel et al. [31] distinguished "direct" replications (attempting to follow the same experimental recipe) and "conceptual" replications ("original methods purposefully altered to test the rigor of the underlying hypothesis", p. 538). In a similar manner, Rosenthal [45] distinguished "fairly precise" and "fairly imprecise" replications. Kelly et al. [26] distinguished four types of replication: (i) Literal replications, where both the measures and the manipulations of an earlier experiment are reused (i.e., both the independent and dependent measures); (ii) operational replications, where the manipulations were replicated but other dependent measures were collected; (iii) instrumental replications, the opposite of operational replications, that is, they replicate measures but change the manipulations; and (iv) constructive replications that involve "an attempt to achieve equivalent results using an entirely original methods recipe" [26, p. 339], hence attempting to ensure external validity by varying both manipulations and measures.

Other papers offer complementary types of replication. Hendrick [15] also distinguished four types of replications: A (a) "strict replication" aims at repeating the study as exactly as possible, focusing on the procedure and context; it matches the literal replication discussed above; (b) a "partial replication" repeats some factors of a study, but introduces deliberate changes to others (covering Kelly's operational and instrumental replications); (c) a "conceptual replication" uses a completely different procedure (as Kelly's constructive replications); and (d) a "systematic replication" combines a strict replication with variations in the conceptual variables. Hendrick [15] argued that a systematic replication can both certify existing findings and potentially extend the scope of their applicability. Tsang and Kwan [48] differentiated six types of replications, defined by the type of population (same data set, same population, or different population) and type of measurement and analysis (same as the original study or different). The main difference to Kelly et al.'s categorization is that Tsang and Kwan consider reanalysis of the original data as a replication, and therefore list "Checking of analysis" and "Reanalysis of data with different methods" as types of replication.

Many studies investigate the extent to which replications are published: The number of replications they find differs widely. Some studies find few or no replications. For example, Sterling [46] found no replications in a sample of 362 articles from psychology journals. In marketing, one paper showed that replication rates were 1.2% [8], whereas a paper covering psychology papers since 1900 found 1.07% replications [31]. Other studies find many replications. Neuliep and Crandall [39] found more than 20% replications in *Journal of Personality and Social Psychology*, and more than 50% of the papers comprised several studies, attempting within a single paper to replicate findings. Some studies have suggested that replications often fail to replicate the original results [48]. These differences in findings may in part reflect differences among fields, in part differences in the definition of replication.

Jones et al. [23] studied replications in the field of human factors. They found 50% to 75% replications by selecting eight papers published in one year of the journal *Human Factors* and looking at citations to those papers in a sixteen-year window. Jones et al. sampled across papers retrieving different number of citations, and found that more highly cited papers were more likely to be replicated. Most of the replications (87% - 94%) were conceptual (in the sense of [15]) and many (19% - 37%) were conducted by the authors of the paper being replicated.

Replications have been much discussed in HCI [e.g., 11,40,51]. Newman [40] argued that much research in HCI present radical solutions, that is, new paradigms to solve problems in interaction. By definition, radical solutions change the problem to be addressed substantially. In contrast, engineering research often enhance existing solutions or methods. Radical solutions do not emphasize existing solutions/methods and do not do incremental refinement; hence, in our reading of Newman, they do not lead to a focus on replicating existing findings. Greenberg and Buxton [11] argued that there is a lack of replications at the ACM CHI conference, claiming reviewer reluctance towards replication studies as one reason. The RepliCHI series of events [51] initiated a discussion of replications in HCI. Among other things, participants in the RepliCHI events have proposed that the CHI conference should create a venue for presenting replications (which happened in 2013). In addition to these general discussions of replications, many influential papers in HCI have included some type of replication [e.g., 7,53].

However, to our knowledge no study has assessed the extent of replications published in the field of HCI or tried to analyze a sample of replications so as to understand their findings relative to the original studies or how replications might be improved.

METHOD

The aims of this study are to (a) discover the extent of replication in HCI research, and (b) characterize the replications done, in particular their type, their results, and the differences to the studies being replicated.

Browsing for Replications

One approach to identifying replications is to search for papers that might contain replications. Unfortunately, no generally accepted terminology for replications exists (authors use “follow up study”, “reproduce results”, etc.), making the selection of appropriate query terms difficult. Neuliep and Crandall [39] pointed out that most replications are not labeled as such; Makel et al. [31] found that only 68% of 500 psychology papers containing the term replication were actually replications. Moreover, it is difficult to use searching to estimate the frequency of replications. At the time of writing, a search on Google Scholar on “human computer interaction” and “replication” returns more than 21,000 entries.

Another approach is to identify a set of papers and investigate how frequently they have been replicated. Jones et al. [23] took that approach in studying how many papers from the 1991 volume of the journal *Human Factors* had been replicated, by whom, and how frequently. This approach has the disadvantages that the sample of papers is crucial to the findings, that replication rates are obtained only for older papers (not current ones), and that the sample size that can be treated is usually small.

A third approach is to browse a set of publication outlets; this was done by for instance Neuliep and Crandall [39]. The drawback of that approach is that replications in HCI most likely are infrequent; studies in other fields typically find low rates. A large number of studies must therefore be examined to characterize replication studies. Nevertheless, we chose to browse publication outlets for replications; this is the preferred method in studies of replications in other fields (e.g., [39]).

We examine four key outlets for research in HCI (see Table 1). We browsed all full papers from the years 2008 to 2010 of a key conference on human-computer interaction and three highly-ranked journals. We had no reasons to expect replications being more frequent in conference proceedings than in journals, so we examined both types of outlet.

Publication outlet	*Papers*	*Eligible*	*%*
ACM Conference on Human Factors in Computing Systems (CHI)	590	265	62
ACM Transactions on Computer-Human Interaction (TOCHI)	63	28	6
Human-Computer Interaction (HCI)	32	22	5
International Journal of Human-Computer Studies (IJHCS)	206	114	27
Total	891	429	100

Table 1. Publications browsed for replications (2008-2010).

Criteria for Eligibility

Before considering whether a paper contains a replication, we checked each paper on certain criteria for eligibility. In particular, a paper must:

- *be empirical*. We do not include literature reviews or papers that theoretically justify or predict particular findings.
- *include quantitative data*. While qualitative studies may build on and relate to earlier findings, such studies typically rely on other notions of overlap among studies than replication [32]. Thus, we restrict our sample to papers that include quantitative data.
- *report an experiment*. This criterion was added for three reasons. First, we wanted to compare the size of effect found in a replication to that of the original study. The notion of effect, and its quantification in effect size, presupposes a comparison of two or more treatments. Second, research may be thought of as descriptive, relational, or experimental [44]. Replications are most likely to be found in experimental research because descriptive and relational research is often exploratory, aiming to describe new phenomena or contexts of use. Third, almost all earlier studies of replications either explicitly or implicitly focus on experiments (e.g., [26]). We wanted to compare our results to these studies.
- *study human interaction with user interfaces*. Although interested in HCI in a broad sense, we did not include papers that simulate computational models or that compare algorithms without data on users’ interaction.

As shown in Table 1 about half of the papers published in the outlets were eligible under these criteria.

What Counts as a Replication?

To code the eligible papers, we fixed a set of operational criteria for determining if a given paper is a replication. In the present paper, a replication is considered an *attempt to confirm, expand, or generalize an earlier study’s findings*. Hence, a replication investigates the reported findings, resulting in either consolidation or potential invalidation or reduction in scope of the original study’s findings. Note that this description of replications does not entail a particular philosophy of science or position on falsification as a scientific ideal.

We operationalize this general formulation as follows.

- A replication must name and reference the original study it replicates; if it did not, it would be very hard to determine if a paper replicated another study. It must also discuss its findings against that earlier study.
- A replication must collect data that relate to the original study; it must express some kind of intent to confirm/expand/generalize that study. We looked for hints about this in the abstract, introduction, method, and discussion sections. The rationale is that if an original study is not mentioned in these sections, it has

not influenced the design or the analysis, and the original study is unlikely to be replicated. In some cases we also analyzed the related work section, for instance when the hypotheses or rationales for a study were developed in that section.

- A replication must concern an original study outside of the paper under consideration; attempts to confirm, expand, or generalize results from earlier parts of a paper are not treated as true replications.
- A replication need not use the word replication or any synonym thereof to be coded as a replication.
- A replication must seek to confirm, expand, or generalize an existing study; using earlier results does not qualify as replication. For instance, a paper may use a questionnaire, validated in an earlier study, to measure a particular construct without being considered replication. Likewise, using a previously established model does not in itself count as a replication (e.g., many studies use Fitts's law to model performance with an input device, but do not attempt to confirm, expand, or generalize that law).
- A replication must seek to confirm, expand, or generalize an existing study; discussing results in relation to earlier work does not make for a replication. For instance, Morris et al. [36] studied SearchBar, a tool for storing, browsing, and querying web search history. In the discussion they noted, "One particularly surprising result was the extensive use of browser tabs among our participants […]. This is in contrast with results reported by Weinreich et al […]" (pp. 1214-1215). We do not consider this paper a replication because the contrast to Weinreich et al. is of minor importance and because the goal of the study was not to compare to the findings by Weinreich et al.

If a paper contains a replication, we distinguish several types based on the literature reviewed earlier (e.g., [26,39]); the types are mutually exclusive:

- *Strict replications* use the same independent and dependent variables as the original study, and attempt to reproduce that study as closely as possible.
- *Partial replications* use deliberate modifications of earlier research, with the aim of testing them in different settings, with different demographic groups of participants, or other operationalizations of variables.
- *Conceptual replications* or *constructive replications* investigate earlier findings, but using different measures, manipulations and settings.

In addition to these types, we include a *number comparison*. This replication type was not seen in the literature, but only in studies that were eligible but did not replicate a finding about differences in means (typically, the difference between means from two experimental conditions) but instead looked at the absolute value for some dependent variable.

We also coded a type of work that is sometimes called replication, but in a much looser sense than the above. A paper with *related experiments* contains several studies that attempt to re-find or challenge each other's findings (i.e., a intra-study replication). For instance, Hoggan et al. [16] investigated adding feedback to touchscreens, reporting two experiments that differed only in the quality of the tactile actuator used to generate the feedback. We do not consider such studies to be replications because they lack the independent scrutiny of earlier research findings that replications offer.

Checking Coding

The above codes were developed over a number of iterations and then applied to papers by the four authors and two research assistants. To test the successfulness of the coding, one rater (the first author) coded a subset of 20% of the papers coded by each of the other raters. Inter-rater agreement using Cohen's kappa indicated substantial to perfect agreement [28]: For eligibility, $\kappa = .9$ (range .7-.99), for replication, $\kappa = .8$ (range .75-.87), and for other codes it was $\kappa = .9$ (range .85-.97).

In addition to the above checks, we also contacted authors from each of the papers that were classified as replications (N = 28). The aim was to let authors input their views. We report outcomes of this coding check in the results section.

Example of Coding

Coding papers was difficult. To give a sense of the complexity involved, we describe some considerations in the coding of a single paper.

Escape [52] is an interactive technique that aims to help users select targets on the touch screens of mobile devices, in particular when targets are small or occluded. It does so by associating a target (say, a pin on a map) with a direction. The user then makes a gesture in the direction associated with a target to select it. The paper describes two experiments; the first compares Escape to an earlier technique called Shift [50]. The first part of the experimental results compare the performance of Shift and Escape; with no other information, this would be an example of building on earlier work (by reusing Shift) but not of replicating a finding. Had the paper for instance repeated the same conditions as in the earlier study of Shift (which included that technique and two baseline conditions) it could have rechecked the earlier findings (for instance, that Shift is faster than unaided touch for small targets). The last part of the results in Yatani et al.'s paper, however, compares directly to Shift by noting [52, p. 290]:

> Although Escape's task time outperformed our reimplementation of Shift in this study, Escape's performance is only marginally better than the original published results […] for targets 12 pixels or less, and is somewhat worse for targets 18 pixels or greater.

Yatani et al. then report a second experiment that compares their implementation of Shift to the original implementation of Shift (see [52], Figure 9 on p. 291). In the terms of the coding scheme presented above, this is a *number comparison*. The second experiment compares some variations on Escape and is not an independent replication (although the paper does contain *multiple studies*). It does not qualify as *related experiments* because the experiment focused on studying new independent variables rather than attempting to re-find the results of the first experiment.

RESULTS

First we discuss how many replications our sample contains and then we turn to the content of those replications.

Frequency of Replication

Among the eligible papers we found 28 papers that contained replications in some form (see Table 2). This is 6.5% of the eligible papers and 3.1% of the full sample. The list of 28 replications can be obtained from the authors; in the present paper we discuss some of them [1,3,5,6,10, 12,13,14,25,29,30,33,38,41,42,47,49,52]. Heer and Bostock [14], for instance, used crowdsourcing to replicate classic results in graphical perception; Pietriga and Appert [42] introduced new types of lenses for magnifying data and compared their performance to earlier work.

Across outlets, the percentage of eligible papers and replications vary. TOCHI and CHI have the lowest number of eligible papers (44% and 45%, respectably), while IJHCS has 57% and HCI has 69%. One outlet contained no replications (TOCHI), while others had more (HCI 5%, IJHCS 6%, CHI 8%) relative to the number of eligible papers. We see no pattern across outlet type to generate eligible papers or papers with replications.

Half of the papers classified as replications explicitly mentioned doing replications, for instance by noting that they "broadly replicate the work of Fitzmaurice and Buxton" [49, p. 2232] or "attempted to replicate results reported by Litman et al." [6, p. 313]. This is more frequent than reported in other fields, where replications are rarely identified as such [39]. Whereas 11 papers did not mention any stem of the word replication, three papers used replication in another sense from that intended here and one paper argued that it did not do a replication [38], even though we classified it as such.

Type	*N*	*%*
Replication	28	7
Strict	3	1
Partial	8	2
Conceptual	17	4
Multiple studies	150	35
Related experiments	67	16
Number comparison	6	1
Eligible papers	429	100

Table 2. Three types of eligible papers. A paper can be simultaneously in the categories Replication, Multiple Studies, and Number Comparison. The subtypes Strict, Partial, and Conceptual are mutually exclusive and exhaustive.

Types of Replication

Table 2 shows the types of replications that were coded; the types attempt to capture the relation between the replication and the initial study. Few *strict replications* were coded (3 of the eligible studies). The paper by Heer and Bostock [14] is one example, as they attempted to do at least two studies that mimicked the setup and measures of earlier studies of graphical perception, changing only the setting in which the study was done (crowdsourced rather than in a lab). Most replications change aspects of the study they replicate, either a little (8, partial replications in Table 2) or much (17, conceptual replications). Nacenta et al. [38] was classified as a *partial replication*. They reused earlier work but adapted implementations, the experimental setup, and the performance measures collected. An example of a *conceptual replication* is the paper by Chen and colleagues [5]. They integrated previous work on two-handed command selection and empirically investigated some benefits of using two hands and of merging commands. Although they compare directly to earlier work (e.g., p. 738), the implementations of interaction techniques and the experimental setup were different.

Papers Reporting Multiple Studies

Among the papers in the sample, 17% contained *multiple studies* (35% of the eligible papers). This number varies among outlets; in particular journals contain more multiple-studies publications (HCI: 44%; TOCHI: 21%; IJHCS: 20%) than conferences (CHI: 14%). Paper length is one explanation of this difference.

Among the papers that contain *multiple studies*, 67 (16% of the eligible papers) were coded as *related experiments.* For instance, Moscovich and Hughes [37] conducted several experiments on multi-touch input. They justified their second experiment by noting: "in the light of Experiment One, we expect that one hand would be better able to coordinate control of an object's position, orientation, and size." (p. 1279). In our analysis, this paper and other papers classified as *related experiments* are conceptually different from replications: while they add credibility and variation to the evidence presented, they are not providing the independent scrutiny possible in replications by independent researchers.

Content of Replications

The content of replications helps answer five questions that we think are of importance to understand how replications are done. We discuss those questions in turn.

Are earlier findings confirmed?

A key purpose in conducting replications is to either confirm earlier findings or show that they do not hold. Among the replications reported, half re-find the effect identified in earlier papers; in one case [6] a study fails to find an effect that an original study had also failed to find.

The remaining half of the replications show mixed results ($N = 7$, 25%), fail to replicate findings ($N = 3$), or could not be coded by all coders ($N = 4$). Showing mixed results and failing to replicate are interesting because they help qualify and scope earlier results. In [38], for instance, the authors noted – after having reported results of a study on how to mitigate difficulties in moving a mouse pointer between screens next to each other – that "the previous findings may seem to contradict the results of the initial evaluation of Mouse Ether" (p. 784). Casiez and Vogel [3] compared to an earlier study and noted that (p. 1714):

> This confirms the results found by Zhai […] while extending it to a wider range of stiffness values. However, Zhai found a 47% difference between the isotonic and isometric rate control devices, which may be explained by the 6 DOF docking task or the different device form factors he used.

How do authors compare to earlier findings?

We were also interested in how authors compared to the original study. Many replication studies (N = 22) used general terms to compare to the original studies. They might write that results are "consistent with" the study being replicated [e.g., 4,13,29], that they "match earlier work" [14], or that they "are similar to those presented by […]" [10]. Some studies argue that they "replicate" earlier findings, without qualifying how or to what extent [e.g., 30]. The use of these general terms means that the comparison to earlier findings is mostly binary, either confirming or failing to confirm.

Before analyzing the sample, we were curious if HCI researchers would use the large literature on going beyond binary outcomes (confirm/non-confirm) toward quantifying effects [44]. Although six papers used effect sizes to report results, only two papers used effect sizes to compare to earlier work. Hartmann and colleagues [13] noted in discussing their results that "the difference in effect sizes between framing without and with exposure to the website is consistent with findings of related previous studies" (p. 862). Another study used a variant of effect sizes (sigma scores) to discuss their results against other work [6].

The frequent binary comparisons used in most replications leave the reader wondering about the data behind the comparisons. Findlater and McGrenere [9] studied adaptive menus and held a hypothesis about the difference between two menus types, called High and Low: "The difference between High and Low would replicate previous findings […]" (p. 1251). They succeed in supporting that hypothesis, but it is not clear how the strengths of the findings compares to the two original studies they replicate. Heer and Bostock [14] replicated earlier work on graphical perception and wrote that they obtained results consistent with earlier work (p. 208). This was supported by a graphical comparison and statistics: "we found a significant effect of plot density (F(3,2415) = 3.49, p = 0.015) but not of background density (F(4,2415) = 0.44, p = 0.779), consistent with Stone and Bartram's findings". However, the effect found by Heer and Bostock is much lower (d = 0.13) than that of original study by Stone and Bartram (d = 1.58 or d = 0.79)[1]. These effect sizes are hard to estimate, but if the difference is as large as the d-values suggest, then the consistency among findings is an apt description at the level of binary significance testing, but not when looking at the actual effects. Both of the studies we have discussed above are exemplary in many respects, but their replication component seems unnecessarily binary.

A few papers compare to earlier work in a non-binary way without using effect size. The techniques for doing so include plotting results against those from original studies [3,10,14] or formulating models of performance [10]. The latter helps comparing across studies, for instance by reducing pointing performance to throughput. Comparisons were sometimes enriched with simple percentages. Chapuis et al. [4] wrote that "Performance results for Point and Bubble cursors are consistent with those in […]", but added "our participants perform faster overall: 10.6% faster for Bubble cursor and 9.6% faster for Point cursor." (p. 1398).

Whose work is replicated?

Most studies replicate the work of other researchers (N = 22). The remainder either replicates their own work (N = 2) or do both (N = 6, e.g., [1]). Note that these cases refer to work reported outside of the replication paper. Although most replications succeed in reproducing earlier studies to the extent that they may be replicated, it is sometimes hard. Some researchers create independent implementations of earlier user interfaces [e.g., 38,52]; in some cases this necessitates extra evaluations to ensure that the new implementation works as in the original study [52].

What counts as a finding?

We had expected that replications attempted to investigate findings from earlier papers and – because we looked for replications in experiments – that those findings would be differences between experimental conditions (e.g., differences in means). This did not always happen.

Six studies do *number comparisons* [e.g., 12,41,50]. Thus, instead of comparing to earlier findings (e.g., differences in means) they compare the numbers from a particular condition directly to numbers in an original study. For instance, Grimes et al. [12] compared the accuracies obtained with a new EEG classification methodology to those from an earlier studies by writing "we were able to replicate classification accuracies reported in previous work (e.g. our accuracies of 92.3% accuracy at 30 seconds is comparable to Gevins' 2-way accuracy of ~95% using 27.5 second windows […])" (p. 841). The key point in *number comparisons* is that the comparison is not made to results of the experimental manipulation (in [12] that would have been task difficulty) but to raw values (in [12] accuracy). Another example is [33] that obtains a correctness rate in

[1] We estimated d = 2*sqrt(df_n*F/df_d) from F-values in the Heer and Bostock paper (those in the quote) and from the Stone and Bartram paper (p. 3, hypotheses 3).

gesture guessing of 46% and compares it to earlier work. *Number comparisons* do not use the experimental component of the paper (which is a necessary to be classified as replication), but instead just checks or relates to a value from the original study. We have not seen this in studies of replications in other fields.

Some studies reuse conditions. To simplify, assume that a replication study builds on an original study that has two conditions, *oldUI* and *controlUI*: presumably, findings that compare these conditions could be interesting to replicate. The replication study may include both of these, and compare them to a new interface, *newUI*. This is done in several cases [e.g., 25,35,41]. In such replications, one could easily check any of the *oldUI* vs. *controlUI* findings, but many studies refrain from doing so. For instance, Karlson and Bederson [25] studied thumb input, building on two earlier interfaces ThumbSpace and Shift. In comparing those interfaces they included a control condition (DirectTocuh), which had been used in the original studies. Karlson and Bederson (study 1) did not in check the earlier findings. With the above discussion of effect size in mind, they could have compared the effect sizes of, for instance, the DirectTouch versus ThumbSpace comparison to the ones in the original study. We consider Karlson and Bederson's study excellent, but from the perspective of replication they could have added even further to our body of knowledge by considering findings relational and by comparing effects, rather than values. In all fairness to Karlson and Bederson, doing such comparisons is hard because earlier papers may not report their data clearly, making the extraction of key statistics hard.

What do authors' of replications think?

As part of this study we contacted authors of studies that had been coded as replications (N = 28). Although we promised not to quote their responses or name individual studies, the patterns in their responses are interesting.

Twenty-three authors responded to our request, representing 82% of the coded replications. Out of those answers, 19 (83%) confirmed our coding, though 13 (57%) reported that while their work contained replication, their study was not planned as a replication. They emphasized that their main goal was not to replicate an original study, but to research new/additional topics and extend the original work substantially.

Four authors (17%) disagreed with our coding. Three authors pointed out that this coding depends heavily on our definition of "replication". They noted differences to the original studies, such as different experimental setup, variables, and research goals. All of the four replications under dispute were coded as partial or conceptual replications, due to the major differences to the original studies. One author in particular disagreed with our coding and mentioned that even if prior work was replicated, the study's main goal was not to replicate anything. Thus, the disagreement with our coding stems from more rigorous definitions of what a replication is than the one we adopted.

The data are in line with other findings reported here. Most replications in HCI are not conducted to replicate earlier findings; typically they seem a byproduct of the necessity to compare new data to prior work.

DISCUSSION

This review of a selection of HCI papers has shown a 3% replication rate. The analysis of replications shows that they mainly confirm earlier findings and that comparisons to earlier studies are often simple. The notion of replication proved difficult, as evidenced by the work taken to code replications and by the authors' comments on papers that we consider replications. Next we discuss these findings and their implications for HCI.

What can we Learn from Replication Studies?

There are two main lessons to be drawn from the 28 replications found in the sample. First, the replications are valuable in that some show that independent research teams can repeat earlier work [49]. In some cases they qualify earlier work [30], suggest that we might not be sure of an earlier finding [38], show that a finding outside of HCI also apply to our field [13], or do an empirical test that explains earlier, seemingly contradictory findings [5]. These examples show that in as far as HCI is an empirical science, replications improve cross-study validity of findings. Although it has been argued that replications are hard because some phenomena in HCI are volatile [27], replications also help see how changing user expectations and technologies affect what we thought we knew.

The second benefit of replications is that they often are specific about what is kept constant and what is changed in empirical work: they motivate *why* a replication is needed. Thereby, critical reasoning about validity issues in earlier work is brought to the fore.

Why so few Replications?

Lack of replications in other fields has shown to be detrimental to proper scientific practice and development of the fields [2,19,48]. Why, then, do HCI researchers not replicate more frequently? Many explanations suggested in other fields probably also hold for HCI, including lack of prestige in replicating, lack of success of replication initiatives, and difficulty in getting replications published. Our analysis has revealed five further points. First, some papers in our sample attempt to improve interaction techniques presented in earlier work. They miss an opportunity, at least in our view, to check if the earlier findings about a technique hold and instead go on to "beat" earlier work with a better technique. This was also evident in the comments from authors, where many explained that they were not trying to replicate, but just to propose new interaction techniques; their focus is on new interaction techniques, not on consolidating earlier work. Thus, building upon earlier studies is widespread, whereas discussing against and replicating them are not.

Second, some papers downplay replication. In [47], Takayama and Nass reported a study that we read as downplaying its replication contribution. They wrote "While this study is similar to Moon's (1998) experiment of computer source proximity in the interviewing context, this paper is the first test of the effects of source proximity in the driving content and the first to investigate effects of source proximity upon physical performance in a safety-critical situation" (p. 176). They might as easily have written "This study replicates the finding of Moon (1998) on computer source proximity in interviewing within a driving context. This provides a first test of the effects of source proximity upon physical performance in a safety-critical situation". Whether that formulation is compatible with the authors' goals is not the point, nor is the specific example the key issue. Mainly, we argue that HCI outlets might be changed in the future, so that authors can more easily write the latter variant and get their work published.

Third, HCI is a multi-disciplinary field, drawing for instance on natural science, social science, and engineering. These fields bring different conceptions of empirical enquiry and of replication to HCI; we already discussed this for qualitative studies. Because of this mix of fields, HCI researchers need both to contribute technology, demonstrate utility to users, and show implications of design. In that mix, replications may end up as a low priority.

Fourth, the distinction between a formative study (e.g., a user study of a new interface) and summative study (e.g., an experiment comparing several variants of interface) might explain the replication rate. Whereas formative studies are useful in driving technological refinements, only summative studies can qualify as a replication. Whereas formative studies change with new technologies and new use contexts, summative studies often use theories and models that abstract from the substantive to the conceptual domain (in the terms of [34]). The low replication rates may be a symptom of priority to empirical studies that evaluate technology, rather than to empirical studies that summarize and challenge conceptual insights.

Fifth, while we emphasize "few" replications, a more positive view may be that HCI is doing well compared to older and more mature fields. We have opted for the rather negative view because of the lack of systematic attempts to replicate work, because of the promise of replications to help emphasize conceptual insights, and because of their utility as a vehicle to clarify, share, and improve experimental work.

Implications and Recommendations

Our papers have some implications for HCI researchers and for editors/chairs in HCI. Note that we are *not* arguing that all studies, or even a majority of studies, should replicate earlier work. We also *do not* want to imply that experiments are a better research method than other methods (see a general discussion in [34] and a HCI-specific discussion in [17]). Rather, our intent is to give a number of recommendations to be taken into serious consideration when designing and reporting experiments for HCI.

How to improve reporting of experiments?

Based on our reading of papers, we think that reports of experiments in HCI, both replications and non-replications, may be improved in several ways:

- Distinguish reuse from replication. Sometimes it is easy to add data analysis to corroborate earlier results.
- Distinguish effect sizes from numbers. What we have called *number comparisons* is to be distinguished from comparing to earlier results (e.g., using effect sizes).
- Include baselines from earlier studies. Instead of just "beating" the best technique from an earlier study, consider including earlier baselines to allow replications of effects.
- Share material from experiments. Some authors had to re-implement interactive techniques and struggled to get the details right. We can help others do that more easily. Also, sharing data allows reanalysis, what some consider a replication [48].
- Describe similarities and differences to earlier work. What is varied and what is kept constant?
- Signal if something is a replication (for instance in the abstract, as in [29]).
- Some papers contain many results, making it hard to identify key comparisons. Compare to the key findings of the work being replicated, not just select findings.
- Discuss against earlier work. All replications discussed against earlier work; in a sample of 102 eligible, non-replication papers from CHI 2010, only 18% had references in the sections following their discussion.
- Authors who write papers containing *multiple studies* (35% of the eligible papers in our sample) may apply the thinking about replications propose here to their intra-paper comparisons, for instance by using effect sizes or by considering the type of replication done.

How to change HCI outlets?

We suggest that editors/chairs rethink their instructions to authors and reviewers. Other fields have seen proposals on how to improve replication rates, such as open science collaboration, journals dedicated to replications, and repositories for replications. For HCI, the guides for submissions for the four outlets we studied could be revised. The instructions for journals, for instance, suggests that "The paper **must be original in some way**: either in its insights, its results, or its approach" (HCI), "… **publishes only original** and significant research papers" (TOCHI), and "…**publishes original research** over the whole spectrum" (IJHCS). The emphases are ours. CHI 2014 authors are told that "it is not just helpful but essential that the submission's **contribution be original**, going beyond any work already reported in other journals or conference proceedings". They can select a contribution type called "Validation and refutation" which mentions replications, though it is not prominent to authors (or reviewers) and seems to contradict the emphasized instructions.

These instructions focus on novelty and originality of research, which works against the publication of replications. The gap between the urge for novelty and adopting replication as a scientific gold standard [21] must be addressed. As discussed earlier, outlets could change their focus from novelty to significance of contributions or, alternatively, just research quality. Incidentally, this approach has been taken by some venues already (e.g., PLOS one). In addition, a dedicated section should mention how the outlet handles submissions with replications. No guideline among the four mentions the word "replication". Increasing the number of prestigious venues that explicitly accept replication studies will lead to more of such contributions in the long run.

The recent discussions in RepliCHI also warrant some comments based on our data. Our key concern is that the definition of replication is less straightforward than some discussions at RepliCHI suggests. The definition of replication was difficult to us as well as the authors we contacted. We applaud the suggestion by John [22] of marking replications at submission time and the CHI 2014 initiative that makes replication as a contribution type (though we prefer them to continue to be part of main track of the conference).

Limitations

This study has a number of limitations that we would like to flag and discuss how to mitigate in future work. First, coding replications is complex. Thus we share with many studies [e.g., 31] the limitation that "if research articles are not framed as replications, then they were not categorized as such" (p. 541). The exclusion of qualitative studies from the sample and of *related experiments* may also be seen as a limitation. Second, we chose to use browsing of publication outlets as a strategy for identifying replications. Third, much more could be done on the dialogue with authors on why they chose to replicate (e.g., on when they do not replicate or on the barriers to replication). Fourth, the coding of replications presupposes much knowledge about the replications' research area. For most papers we lack that, perhaps over-emphasizing what authors have written.

CONCLUSION

We have studied replications in HCI through browsing papers from journals and publications. Three percent replications were identified and analyzed. We have suggested several implications of the analysis for HCI outlets that could increase replication rates.

ACKNOWLEDGMENTS

We are grateful to Malene Jacobsen and Mathias Andersen for help in coding. We also thank the many authors that answered our questions about their work carefully and patiently. Sebastian Boring and Antti Oulasvirta commented on the draft, which helped us immensely.

REFERENCES

1. Appert, C., Chapuis, O., and Pietriga, E. High-precision magnification lenses. In *Proc. CHI* (2010), 273–282.
2. Begley, C.G. and Ellis, L.M. Drug development: Raise standards for preclinical cancer research. *Nature 483*, 7391 (2012), 531–533.
3. Casiez, G. and Vogel, D. The effect of spring stiffness and control gain with an elastic rate control pointing device. In *Proc. CHI*, (2008), 1709–1718.
4. Chapuis, O., Labrune, J.-B., and Pietriga, E. DynaSpot: speed-dependent area cursor. In *Proc. CHI* (2009), 1391–1400.
5. Chen, N.Y., Guimbretière, F., and Löckenhoff, C.E. Relative role of merging and two-handed operation on command selection speed. *International Journal of Human-Computer Studies 66*, 10 (2008), 729–740.
6. D'Mello, S.K., Graesser, A., and King, B. Toward Spoken Human–Computer Tutorial Dialogues. *Human–Computer Interaction 25*, 4 (2010), 289–323.
7. Egan, D.E., Remde, J.R., Gomez, L.M., Landauer, T.K., Eberhardt, J., and Lochbaum, C.C. Formative design evaluation of superbook. *ACM Transactions on Information Systems (TOIS) 7*, 1 (1989), 30–57.
8. Evanschitzky, H., Baumgarth, C., Hubbard, R., and Armstrong, J.S. Replication research's disturbing trend. *Journal of Business Research 60*, 4 (2007), 411–415.
9. Findlater, L. and McGrenere, J. Impact of screen size on performance, awareness, and user satisfaction with adaptive graphical user interfaces. In *Proc. CHI* (2008), 1247–1256.
10. Forlines, C. and Balakrishnan, R. Evaluating tactile feedback and direct vs. indirect stylus input in pointing and crossing selection tasks. In *Proc. CHI,* (2008), 1563–1572.
11. Greenberg, S. and Buxton, B. Usability evaluation considered harmful (some of the time). In *Proc. CHI* (2008), 111–120.
12. Grimes, D., Tan, D.S., Hudson, S.E., Shenoy, P., and Rao, R.P. Feasibility and pragmatics of classifying working memory load with an electroencephalograph. In *Proc. CHI,* (2008), 835–844.
13. Hartmann, J., De Angeli, A., and Sutcliffe, A. Framing the user experience: information biases on website quality judgement. In *Proc. CHI* (2008), 855–864.
14. Heer, J. and Bostock, M. Crowdsourcing graphical perception: using mechanical turk to assess visualization design. In *Proc. CHI* (2010), 203–212.
15. Hendrick, C. Replications, strict replications, and conceptual replications: Are they important? *Journal of Social Behavior & Personality*, (1990).
16. Hoggan, E., Brewster, S.A., and Johnston, J. Investigating the effectiveness of tactile feedback for mobile touchscreens. In *Proc. CHI* (2008), 1573–1582.

17. Hornbæk, K. Some Whys and Hows of Experiments in Human–Computer Interaction. *Foundations and Trends in Human–Computer Interaction 5*, 4 (2013), 299–373.
18. Huizenga, J.R. *Cold fusion: The scientific fiasco of the century* (1992).
19. Hunter, J.E. and Schmidt, F.L. *Methods of meta-analysis*. Sage, 2004.
20. Ioannidis, J.P. Why most published research findings are false. *PLoS medicine 2*, 8 (2005), e124.
21. Jasny, B.R., Chin, G., Chong, L., and Vignieri, S. Again, and again, and again…. *Science 334*, 6060 (2011), 1225–1225.
22. John, B. Avoiding "It's JUST a Replication", *RepliCHI 2013 Workshop*.
23. Jones, K.S., Derby, P.L., and Schmidlin, E.A. An investigation of the prevalence of replication research in human factors. *Human Factors, 52*, 5 (2010), 586–595.
24. Kaplan, A. *The Conduct of Inquiry: methodology for behavioral science*. Chandler, San Francisco, CA, 1964.
25. Karlson, A.K. and Bederson, B.B. One-handed touchscreen input for legacy applications. In *Proc. CHI* (2008), 1399–1408.
26. Kelly, C.W., Chase, L.J., and Tucker, R.K. Replication in experimental communication research: An analysis. *Human Communication Research 5*, 4 (1979), 338–342.
27. Lallemand, C., Koenig, V., and Gronier, G. Replicating an International Survey on User Experience: Challenges, Successes and Limitations. *RepliCHI 2013 Workshop*.
28. Landis, J.R. and Koch, G.G. The measurement of observer agreement for categorical data. *Biometrics*, (1977), 159–174.
29. Lee, E.-J. Flattery may get computers somewhere, sometimes: The moderating role of output modality, computer gender, and user gender. *International Journal of Human-Computer Studies 66*, 11 (2008), 789–800.
30. Lee, E.-J. I like you, but I won't listen to you: Effects of rationality on affective and behavioral responses to computers that flatter. *International Journal of Human-Computer Studies 67*, 8 (2009), 628–638.
31. Makel, M.C., Plucker, J.A., and Hegarty, B. Replications in Psychology Research How Often Do They Really Occur? *Perspectives on Psychological Science 7*, 6 (2012), 537–542.
32. Marshall, C. and Rossman, G. B.. *Designing qualitative research,* Sage (1999).
33. Martin, B. and Isokoski, P. EdgeWrite with integrated corner sequence help. In *Proc. CHI* (2008), 583–592.
34. McGrath, J.E. Methodology matters: Doing research in the behavioral and social sciences, in *Human-computer interaction* (1995), Morgan Kaufmann, 152-169.
35. Moffatt, K. and McGrenere, J. Steadied-bubbles: combining techniques to address pen-based pointing errors for younger and older adults. In *Proc. CHI* (2010), 1125–1134.
36. Morris, D., Ringel Morris, M., and Venolia, G. SearchBar: a search-centric web history for task resumption and information re-finding. In *Proc. CHI* (2008), 1207–1216.
37. Moscovich, T. and Hughes, J.F. Indirect mappings of multi-touch input using one and two hands. In *Proc. CHI* (2008), 1275–1284.
38. Nacenta, M.A., Mandryk, R.L., and Gutwin, C. Targeting across displayless space. In *Proc. CHI* (2008), 777–786.
39. Neuliep, J.W. and Crandall, R. Everyone was wrong: There are lots of replications out there. *Journal of Social Behavior & Personality*, (1993).
40. Newman, W. A preliminary analysis of the products of HCI research, using pro forma abstracts. In *Proc. CHI* (1994), 278–284.
41. Olwal, A., Feiner, S., and Heyman, S. Rubbing and tapping for precise and rapid selection on touch-screen displays. In *Proc. CHI* (2008), 295–304.
42. Pietriga, E. and Appert, C. Sigma lenses: focus-context transitions combining space, time and translucence. In *Proc. CHI* (2008), 1343–1352.
43. Reaves, M.L., Sinha, S., Rabinowitz, J.D., Kruglyak, L., and Redfield, R.J. Absence of detectable arsenate in DNA from arsenate-grown GFAJ-1 cells. *Science 337*, 6093 (2012), 470–473.
44. Rosenthal, R. and Rosnow, R. *Essentials of behavioural research.* McGraw-Hill (1991).
45. Rosenthal, R. Replication in behavioral research. *Journal of Social Behavior & Personality*, 5, 1990, 1-30.
46. Sterling, T.D. Publication decisions and their possible effects on inferences drawn from tests of significance—or vice versa. *Journal of the American statistical association 54*, 285 (1959), 30–34.
47. Takayama, L. and Nass, C. Driver safety and information from afar: An experimental driving simulator study of wireless vs. in-car information services. *International Journal of Human-Computer Studies 66*, 3 (2008), 173–184.
48. Tsang, E.W. and Kwan, K.-M. Replication and theory development in organizational science. *Academy of Management Review 24*, 4 (1999), 759–780.
49. Tuddenham, P., Kirk, D., and Izadi, S. Graspables revisited. In *Proc. CHI* (2010), 2223–2232.
50. Vogel, D. and Baudisch, P. Shift: a technique for operating pen-based interfaces using touch. In *Proc. CHI* (2007), 657–666.
51. Wilson, M.L., Mackay, W., Chi, E., Bernstein, M., Russell, D., and Thimbleby, H. RepliCHI-CHI should be replicating and validating results more. In *Extended Abstracts of CHI* (2011), 463–466.
52. Yatani, K., Partridge, K., Bern, M., and Newman, M.W. Escape: a target selection technique using visually-cued gestures. In *Proc. CHI* (2008), 285–294.
53. Zhai, S., Conversy, S., Beaudouin-Lafon, M., and Guiard, Y. Human on-line response to target expansion. In *Proc. CHI* (2003), 177–184.