Replication in Computing Education Research: Researcher Attitudes and Experiences

Alireza Ahadi
Univ. of Technology, Sydney
Australia
alireza.ahadi@uts.edu.au

Arto Hellas
University of Helsinki
Finland
arto.hellas@helsinki.fi

Petri Ihantola
Tampere Univ. of Technology
Finland
petri.ihantola@tut.fi

Ari Korhonen
Aalto University
Finland
ari.korhonen@aalto.fi

Andrew Petersen
Univ. of Toronto Mississauga
Canada
petersen@cs.toronto.edu

ABSTRACT

Replicability is a core principle of the scientific method. However, several scientific disciplines have suffered crises in confidence caused, in large part, by attitudes toward replication. This work reports on the value the computing education research community associates with studies that aim to replicate, reproduce or repeat earlier research. The results were obtained from a survey of 73 computing education researchers.

An analysis of the responses confirms that researchers in our field hold many of the same biases as those in other fields experiencing a crisis in replication. In particular, researchers agree that original works – novel works that report new phenomena – have more impact and are more prestigious. They also agree that originality is an important criteria for accepting a paper, making such work more likely to be published. Furthermore, while the respondents agree that published work should be verifiable, they doubt this standard is widely met in the computing education field and are not eager to perform the work of verifying others' work themselves.

CCS Concepts

•General and reference → Surveys and overviews; •Social and professional topics → Computing education; •Applied computing → Education;

Keywords

research process; replication; reproduction; verification; validation; publication bias; computing education research

1. INTRODUCTION

Building knowledge – conducting research – is not easy. Significant research effort is invested into projects that fail to yield results or, more commonly, *publishable* results, which leads to manuscripts occupying space in a virtual file drawer. At other times, the investment in time and effort pays off, leading to a publication. However, in some published articles, the results being reported are the result of unidentified or implicit factors. These results are examples of phenomena that are not robust: they may not be observable in a study organized in another context. This is especially common in social sciences such as education, where researchers may find it challenging to replicate even their own work.

Our work stems from the observation that while published studies in computing education are expected to build on previous work, only a small number seek, explicitly, to replicate or reproduce existing work. Why is this so? New results are abundant, but why are they only rarely re-tested and validated? Do researchers believe that replicating a study conducted by others will not generate sufficient prestige? Is novelty the key factor? Do researchers consider existing results difficult or impossible to replicate?

In fields such as medicine and psychology, discussions of the importance of replication have stemmed from observations that published results do not hold up under further scrutiny [20, 34]. As a result, Schooler, among others, recently initiated a movement to increase the replicability of findings by having a set of labs replicate each others' findings before publication [43]. As a discipline, the computing education research community is still far from such a movement, but our hope is to increase awareness of the importance of replication, the need to revisit accepted theories and studies, and the need to build upon tested and validated work.

Here, we explore results from a survey of 73 computing education researchers. The survey solicits the perspective of the community on studies that seek to replicate, reproduce or repeat some earlier research. We investigate the *reasons why researchers have (or have not) sought to replicate past research and aim to determine the reasons for and against conducting replication studies. Opinions on the value and publishability of replication studies are also solicited.*

This article is organized as follows. First, in Section 2, we review related streams of research where the focus has been on the replication of research results. Next, we describe our research methodology in Section 3. Results from the survey are presented in Section 4 and discussed in Section 5. Our summary of the responses and suggested responses are presented in Section 6.

2. RELATED WORK

Various groups have sought to differentiate between different forms of replication study. For example, Lykken proposed that replications could be literal (exact duplications to the point of sampling decisions), operational (where experimental or sampling procedures are duplicated), or constructive (where the original methods are explicitly avoided but the goal is to provide evidence of the phenomenon being studied) [24]. Asendorpf et al. define the concepts of reproducibility (using the same data set), replicability (using new data in the same experimental space), and generalizability (confirmation that an experimental space is not impacted by unmeasured variables) [4]. A handful of similar discussions exist in different fields [16,18]. Schmidt presented a comprehensive summary of the issues that contributed to varying definitions of reproducibility in 2009 [42].

In this paper, we do not distinguish between the terms "reproduction" and "replication." Instead, we focus on perceptions of the value of projects that seek to replicate, in any way, previously published research, and we use the term "replication" to describe any such activity.

2.1 Replication in non-CS Disciplines

The sciences and social sciences are currently suffering from a crisis in confidence, and a lack of faith in the replicability of published results is at the core of this issue. Replicability is a critical element of the scientific method [35], yet recent studies have suggested that a significant fraction of published results cannot be replicated. For example, a recent study in psychology claimed that only 40% of studies were replicable [10]. This result ignited concerns about flawed attempts at replication [14] and questionable research practices [3].

Recently, a multi-disciplinary survey administered by *Nature* found that "73% [of respondents] said that they think that at least half of the papers in their field can be trusted," meaning that over a quarter of respondents trusted less than half of the publications in their field [5]. This reflects a more general concern that researchers in a wide range of disciplines have expressed about the state of replication in their fields (e.g., biology [7], health care [19], political science [15], and computational science [37]).

Concerns about replicability are supported by the observation that replication studies are rare [27,35] and, in some fields, have become more rare over time [12]. Various authors have drawn connections between the rarity of replications and the importance accorded to innovation and original research [4,25] or to the perception that replications are not publishable at journals [13,46]. The *Nature* survey found that a minority of researchers had even tried to publish a replication, though, in a positive note, 24% had published a successful replication and 13% had published a failure to replicate a result [5]. This is balanced by a roughly equivalent number (10%) of respondents who reported that they were unable to publish a failed replication.

Various fields have responded to the finding that key results are not replicable by attempting to improve the state of science in their discipline. Suggestions include calls to adopt more careful experimental procedures, to increase reporting of experimental design decisions, to require publication of data and research materials, and even to change faculty promotion criteria to favor "quality" results and to reward attempts to replicate results [4,21].

2.2 Replicability in Computing Education

Replicability is particularly difficult in social sciences such as education, and less than a percent of studies published in educational journals are replications [26]. Part of the issue lies in the nature of educational research. Exact replications may not be possible in all situations, but in education (and the social sciences in general), at least some phenomena should be replicable [25,42].

At the same time, perceptions of the value – and impact – of replications also contribute to the dearth of published replication studies. For example, an editor of *Educational Psychology* was quoted as saying that he "doesn't publish replications studies 'unless they cover new ground'" [6]. More recently, Mark Guzdial noted that a recent attempt to publish a replication of an important instrument for measuring introductory computer science knowledge met initial resistance "because replicating the FCS1 wasn't deemed to be as noteworthy as the original work" and encountered additional resistance after resubmission due to concerns that the new work was too similar to the work it replicated [17].

As in other fields, researchers in computer science education have raised concerns about the quality of research in the discipline [9] and, more specifically, the absence of replication. Al-Zubidy et al. raise a particularly significant concern: they explored the use of rigorous methods and evidence at SIGCSE and suggested that the lack of replication in the computing education domain inhibits the maturation of the field [2]. The de-emphasis on replication makes it impossible to conduct the meta-analyses that lead to the generation of theories. Al-Zubidy et al.'s work [2] is particularly notable as being a replication of prior work completed by Valentine [49] and extended by Randolph et al. [40].

Notably, Randolph and Bednarik found that authors emphasized statistically significant results and de-emphasized non-significant results, which suggests a need for replication studies to identify potentially non-generalizable results [41]. Others have called for improving transparency and increasing the amount of detail in publications, so as to facilitate replications [18] and for more recognition of the value of replication studies [17].

2.3 Examples of Replication in Computing Education Research

The computing education literature does contain a number of studies that at least partially replicate previous works. As an example, a search for "Education" in the CCS concepts and "replication" in author reported keywords resulted in 6 hits in the ACM Digital Library [18,31,32,36,47,48]. If the search is relaxed to articles that contain the word "replication" in any field, a total of 91 articles are found. While far from a complete search, this number is a disappointingly small fraction of the 21,706 articles with the term "Education" in the CCS concepts in the ACM Digital Library[1].

The "replication" examples found take several different forms. Several are ITiCSE working group reports [18,23,48], where researchers collaboratively target a particular topic or a problem. A number of studies analyze and extend previous studies on instructional techniques such as peer instruction [39] by considering a a new educational context. We also observed several examples of studies which revisit the topic of a popular article, with the new work including a

[1] Queries performed in August 2016.

partial replication and extension of the work [22, 32, 33, 50]. A handful of studies explicitly replicated past quantitative analyses using data from new contexts [1, 38].

3. METHODOLOGY

The data presented in this study was generated using a survey targeting researchers active in computer science education. The goal of the survey, which is reproduced in Appendix A, is to gather perspectives on, first, the perceived value of replication studies to researchers and the community and, second, the relative difficulty of publishing such studies. The survey contains 38 items organized in four parts: demographics, beliefs about replication in computing education research, experience with replication studies, and the a question about the definition of the term "replication."

Most of the items on the survey relating to the value of replication studies to researchers were based on common beliefs about requirements for faculty promotion or were based on attributes identified as necessary for science. Items in the survey relating to the relative difficulty of publishing replication studies were derived from a survey on attitudes toward replication completed in the natural sciences [5]. The survey was trialed by colleagues of the authors. Items which produced highly variant responses in the trial were reviewed, and the results of the review and feedback from the respondents trialing the survey were used to update the questions.

The survey contained both quantitative data (yes/no responses and Likert-scale data) and text responses to free response questions. The quantitative data was aggregated, and aggregate statistics are reported. Correlations between questions were also calculated to identify potential relationships between question pairs. Bonferroni correction [11] was used to counteract the problem of performing a large number of statistical tests. The qualitative data was coded to form categories, and the categories and counts in each category are reported. Particularly illustrative individual quotes from the free response questions were identified and included in the text.

3.1 Demographics

The final survey was widely advertised using the SIGCSE-members mailing list, reddit, and the personal networks of the authors (including university mailing lists, twitter, and facebook). 79 responses were received. 6 responses were largely incomplete or were duplicates, leaving 73 responses in the set that were analyzed.

Table 1: Demographic data for survey respondents

Locations		Experience	
Africa	1	0-5 years	10
Asia	3	6-10 years	18
Europe	13	11-15 years	11
North America	49	15-20 years	9
Oceania	7	20-30 years	19
South America	0	30+ years	6

Table 1 summarizes the demographic data (geographic location and years of experience) collected from respondents of the survey. A significant majority of the respondents spent the majority of their career in North America, with Europe and Oceania relatively well represented. Very few responses were received from other regions.

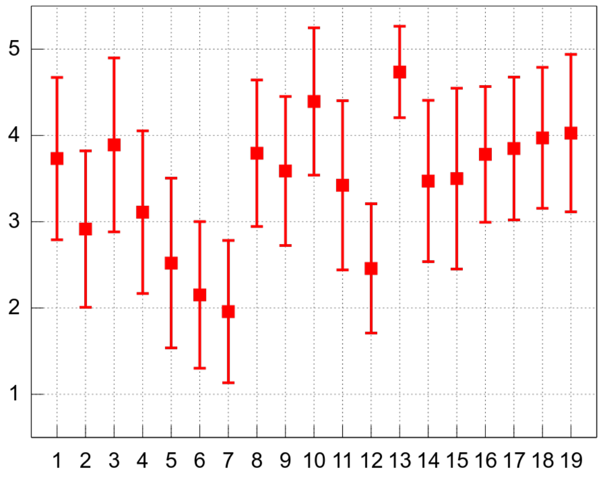


Figure 1: Averages and Standard Deviations of the Responses to the Survey Questions (Appendix A, part 2). Averages are marked with a square, and the Standard Deviation (+-) with a bar. Responses to survey items were given using a five-point Likert-scale, where 1 corresponds to strongly disagree and 5 corresponds to strongly agree.

The respondents were well distributed in terms of years of experience teaching or researching at the university level. The vast majority (63) of respondents were university teachers or researchers, with seven reporting their status as "Ph.D. student" and three as "Other."

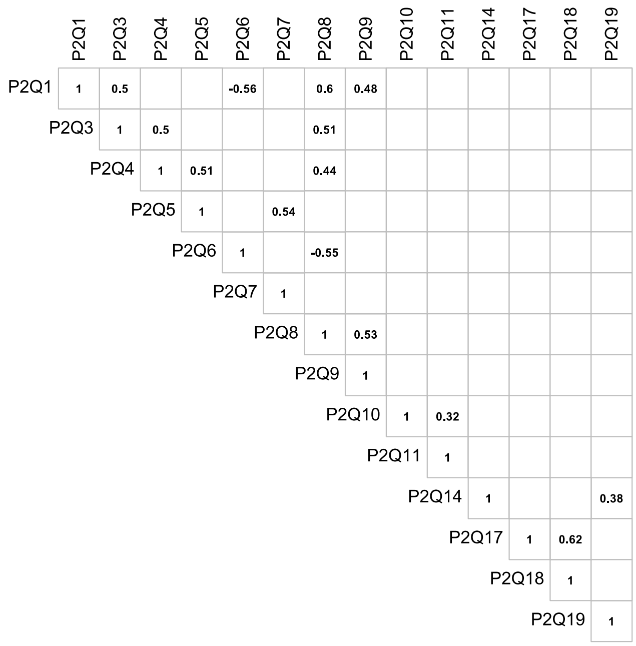
The vast majority of respondents (63) self-reported that they are active in the CS Education research community. (We chose not to define "activity" in the community, so as to include anyone who felt they have a role in the community.) Most other respondents (6) are researchers in other computing fields, though a small number (3) are active in computing education but declined to describe their work as research and one respondent declined to provide detail. Most respondents (66) teach. Those who do not reported being active in computing education research.

4. RESULTS

This section is organized to match the structure of the survey (in Appendix A). The demographic data in Part 1 was presented in the previous section (3.1). Section 4.1 presents the quantitative results drawn from the Likert questions in Part 2 of the survey. The numeric responses in Part 3, which relate to respondent's confidence in the replicability of published work and their experience publishing replication studies, are analyzed in Section 4.2. The long form responses explaining why they choose to engage (or not) in replications are summarized in Section 4.3). Finally, Section 4.4 explores how the community differentiates between different forms of replication.

4.1 Replication and Computing Education Research

Figure 1 presents the results of the questions in Part 2 of the survey (P2Q1–P2Q19), which is related to beliefs and perceived attitudes toward replication in computer science



	P2Q1	P2Q3	P2Q4	P2Q5	P2Q6	P2Q7	P2Q8	P2Q9	P2Q10	P2Q11	P2Q14	P2Q17	P2Q18	P2Q19
P2Q1	1	0.5			-0.56		0.6	0.48						
P2Q3		1	0.5				0.51							
P2Q4			1	0.51			0.44							
P2Q5				1		0.54								
P2Q6					1		-0.55							
P2Q7						1								
P2Q8							1	0.53						
P2Q9								1						
P2Q10									1	0.32				
P2Q11										1				
P2Q14											1			0.38
P2Q17												1	0.62	
P2Q18													1	
P2Q19														1

Figure 2: Correlation plot including only the pairs where statistically significant correlation was observed. As 210 pairs (21 variables) were tested, the Bonferroni corrected p-value used as the threshold is 0.05/210.

education. Relatively few items were contentious; generally, the respondents were in agreement on the survey items. P2Q2, regarding the relative difficulty of replication and original studies, and P2Q4, which raises the question of whether replications are worth completing given the reward, stand out as being controversial.

Several of the questions in part two of the survey were highly correlated. We calculated Spearman correlations to all pairs of ordinal (i.e., Likert-scale) and interval (i.e., age) variables in the second part of the survey as well as P3Q1, which was also numerical. The statistically significant correlations (after Bonferroni correction) are shown in Figure 2. A further manual analysis of the correlations between the variables led into the following connections:

- *Difficulty of publishing (P2Q1, P2Q3, P2Q6, P2Q8, P2Q9):* The relative difficulty of publishing replication studies and original findings were negatively correlated, with general agreement that replication studies are more difficult to publish. Moreover, the relative difficulty of publishing replication studies was moderately correlated with the perceived prestige of original studies studies as well as the value of publishing in general.

- *Value of publishing replications (P2Q3, P2Q4, P2Q5, P2Q7, P2Q8):* The perceived value of publishing original studies was positively correlated with the perception that replication studies are less valuable in terms of obtaining citations and grant funding. Moreover, there was a positive correlation between the concern that replication studies take resources from original work and the belief that replication studies are mechanical exercises.

- *Promotion criteria and difficulty of publishing (P2Q10, P2Q11, P2Q14, P2Q19):* There was agreement that researchers should publish large amounts of papers and that more of these articles should be original work, rather than replication studies. Somewhat inconsistently, respondents who agree that a high citation count is important are also more likely to state that work submitted for publication should be grounded in previously published results.

- *Importance of replicability and verifiability (P2Q17, P2Q18):* Agreement on the importance of replicability and verifiability during the evaluation process was also identified, with respondents who were inclined to see replicability as important also marking verifiability as important.

4.2 Replication Experiences

Part 3 of the survey asked participants to reflect on their experiences reviewing and attempting to replicate published work. The results suggest considerable pessimism about our discipline's ability to replicate existing work.

In particular, P3Q1 asks what proportion of published results in computing education are replicable. 11 participants declined to respond to the question, but the maximum value reported, by two participants, was 70%. The mean, across the 62 respondents who provided a figure, was 34.2% with a standard deviation of 17.1. The median was 30%. In comparison, in the cross-disciplinary survey completed by *Nature*, the *average* on this question was higher, for several hard science disciplines, than our maximum.

This perception may be driven by experience with replication studies. 35 respondents (48%) reported that they had failed to replicate one of their own results, and (40%) reported attempting and failing to replicate the result of another researcher.

Publication rates for replications, both successful and unsuccessful, are relatively low and reflect the low number of replications currently published. 7 (10%) were able to publish a successful replication, and 6 (8%) published a failed replication. Similar numbers – 8 (11%) and 6 (8%) – failed to publish successful and unsuccessful replications, respectively. From the phrasing of the question, however, we cannot determine if those replication projects were eventually published (after one or more rejections) or were never successfully published.

4.3 Reasons for (not) Conducting Replication Studies

P3Q8 (15 responses) and P3Q9 (18 responses) asked respondents to explain why they chose to *conduct* or *not conduct* replication studies, respectively. 13 participants declined to respond to either question. Although we asked for responses to only one or the other of these, 27 respondents answered both questions. Two authors categorized the responses of participants who answered both questions, adding them to P3Q8, P3Q9, or both. (e.g., if the response to P3Q9 was "see the other response," we omitted that entry and added the participant's answer to P3Q8 to the analysis.) In 14 cases, the responses were moved to either P3Q8 or P3Q9, but not both. In 13 cases, the respondent provided substantial answers to both prompts, so their responses were included in both P3Q8 and P3Q9. In total, we obtained 32 responses to P3Q8 and 41 to P3Q9.

4.3.1 Reasons to Conduct Replication Studies

Six of the 32 free responses to P3Q8 explained why and what kind of replication studies they have conducted. For example, "to ensure a result is valid", "...expand a previous study...", or "...continued replicating the study until [they] were successful."

However, most of the respondents explained their motivation for completing replications. Eight respondents explicitly used the words "important" or "essential" to describe the role of replications in the scientific progress. Eleven mention the need for confirmation or falsification: "to confirm findings of others", "to verify earlier work and to rule out the effects of site-specific factors, or to elucidate which other factors may be relevant", "to validate others' research", etc.

Many of the responses that mentioned the importance of confirmation use the verb "validate." For example, "I did not believe the result and replicated the study to validate the results." This kind of suspicion towards published results was mentioned in 7 responses: "See if a surprising result really works out, or was a fluke," "Replication ... provides confidence in previously published results," and even, "We have published results that we didn't believe ourselves, so we tried to replicate them."

Many respondents have also realized that this kind of validation process can have two outcomes. One can confirm a previous study, i.e., "To verify previous results and to build understanding on what are the context-specific factors that contribute to study outcomes." Or, one can fail to find evidence for a study. This experience is not always a positive or effective one:

> I found it very difficult to accept the published result, and felt that too much was left unspecified; so I persuaded the original authors to work with me to replicate the work more rigorously. As a consequence, they retracted their original claims. Unfortunately, the original claims, which were sensational, still attract far more attention than the replication.

Finally, seven responses mention that replications enable extensions that might themselves be novel (and hence, valued) contributions: "Replication is the first step in building upon someone else's work", and "Replication studies are a good starting point for original work that builds on previous findings."

4.3.2 Reasons to Not Conduct Replication Studies

For the 41 responses to P3Q9, we found 5 categories concerned about the *value* of conducting replication studies (VAL: 12 occurrences), the probability that the study will be *accepted for publication* (ACC: 7), a *preference for novel* work (NEW: 12), a *lack of opportunities* (OPP: 10), and *inability* to replicate due to lack of detail in the original publication (INA: 6). One response did not fit in any of these. Later, we combined the VAL and ACC categories as we considered all ACC responses to belong to VAL as well. The other four categories remain distinct.

The largest category (VAL) reflects concerns about the value of replication studies. For example, on respondent noted a, "lack of prestige; poor investment of time and energy in a system that seems to value replication studies very little." More than half of this category is composed of respondents who are concerned with the perceived difficulty of getting replication studies published (ACC): "I haven't conducted a replication study that I intended to publish, in part because I have the sense that they are difficult to publish." Another respondent is more cynical: "Because the community doesn't care and it wouldn't get accepted [...]"

We believe this fear is well justified as one respondent is a reviewer that suggests that they would reject papers that are pure replications:

> As a reviewer I would expect some new insights even if the findings are similar to the original work, new interpretations, new conclusions, new questions, some kind of a eureka. So I think there are many times more replication studies conducted than published.

The second category, NEW, reflects a preference for engaging in novel projects. These respondents argued, for example, that, "Originality seems to be more interesting to the researcher herself, to the community in general and to your funders." or that "[I am] More interested in doing new things." Some respondents in this category appear to not consider replications to be valuable work: "I prefer to do original research [only]."

The other two categories are smaller. Category OPP includes respondents that indicate they have had no opportunity to perform a replication. Many of these responses also note a problem with the value of replication work: "I have not as yet, but am aiming to. I undertake studies because they are interesting, not because they are publishable." In contrast, respondents in category INA claim not that they haven't had an opportunity but that, in general, it's *not possible* to perform replications: "Articles often in our field do not provide sufficient information for replication" and "Replication information is very hard to come by in CS education. Often methods or materials are not clearly specified or cited."

Even if we accept that OPP and INA reflect a non-negative attitude to replications, 24 respondents – almost a third of the 73 total respondents – do not see value in replicating previous studies.

4.4 Types of Replication Studies

Survey participants were also asked about the definition of the term "replication" and whether there are differences between "studies that aim to replicate, reproduce or repeat some earlier research?" 23 participants declined to answer the question, leaving 50 respondents.

Most of the remaining respondents – 31, or 62% – saw no difference between the terms. Of the remaining 19 respondents, nearly half (8) said "yes" with no further detail, and the other respondents were roughly split into two groups.

Four respondents suggested that some of the terms refer to studies intended to produce the same results while other terms refer to studies where the intention is to extend the work. However, there is no consensus on which term refers to work that is intended to extend the original study:

> I have not looked at these studies as a group. I can imagine two: (1) a study that seeks to copy and verify the original results (2) a study that seeks both to do that and to explore things further.

The other respondents provided a definition that differentiates between studies that focus on reanalysis or verification of the initial experimental setup (either "repeat" or "replication," depending on the respondent) and studies that study the original phenomenon under different conditions (a "reproduction"):

> Yes! Repeat is the mechanical replication (same researchers, same method, maybe even same data and analysis that hopefully give the same results unless there is a systematic error somewhere in the analysis), and reproduction is the other extreme where everything (researchers & method that implies also different analysis method, but maybe give similar results and conclusions) has changed compared with the original study except the research questions. Replication is a generic term that can be anything between these two extremes.

These responses suggest that there are few opinions – and even less agreement – in computer science education on the terms used to describe studies that attempt to replicate previous findings. At the least, this suggests a lack of experience with the issues involved in replicating previous work and may also suggest a lack of interest.

5. DISCUSSION

Many of the opinions expressed in the survey of computer science education researchers reflect concerns that have been expressed in other fields. In particular, the respondents generally agreed with P2Q3 (original studies are more prestigious than replication studies) and P2Q16 (originality is an important criteria for publication), suggesting that there is a general bias toward original work. This corroborates Guzdial's anecdote about the difficulty of publishing a replication [17].

Other factors combine with this bias toward original work to explain why relatively few replication studies are published in computer science education. Respondents who place more value on volume of publications for promotion also believe that original findings are easier to publish, suggesting that researchers in the field should focus on original work prior to promotion in order to maximize their chances of publishing at the volume expected by their peers. In addition, while replication studies might matter for promotion (P2Q12), new faculty are expected to focus on original work (P2Q11). Furthermore, since there is agreement that replications are generally worth less, both in terms of citations and funding, this preference for publishing original work is likely to persist later in faculty careers.

This general preference to avoid replication work does not align with beliefs about the qualities published work should have. There is only weak agreement that a researcher's work should be validated for them to be promoted (P2Q15), but respondents more strongly agree that a researcher's work should be reproducible or verifiable (P2Q17 / P2Q18). Our community values the opportunity to validate work but is unwilling to commit to requiring that work or to perform that work themselves.

As we can see from the responses in Part 3, the community does agree that the ability to replicate work is important, but it is not clear that this standard is currently being met. A significant amount of authors (and potential reviewers) do not see the value of a replication of a previous study if no new findings are reported. In addition, there are even more authors that "prefer to do original research" (regardless of whether they see the value of such work). This might be because many CS education researchers do not see replications as significant contributions to the field in the same way, for example, that HCI does. Furthermore, it is unclear who is available to perform replications, as researchers at all levels are incentivized to prioritize higher impact, original work.

5.1 Replication is Difficult

One reason for not conducting replications is that they are difficult. This may be because of incomplete reporting of the original research [18] or because of the highly contextualized nature of the learning sciences. As one respondent wrote:

> My estimate of 30% for how much of published CS Ed work is reproducible is not meant as a comment on the quality of CS Ed work. It's that so much of our work is highly-contextualized (specific classes, specific schools, specific classes of students) that it would be hard to reproduce exactly, and if we change the context, I'd bet that many of the results would change, too.

Insights like the above are crucial reasons to value replication work. Each context has multiple factors that may influence the study outcomes, and only through rigorous study that includes replication can we start to identify those factors. Recent work on the traditional rainfall problem which identified several possible confounding factors is an example of the importance of replication studies to the process of identifying contextual factors [44].

It is also possible that replicability depends on the purpose of the original research – whether it is, for example, descriptive, evaluative, or formulative [30]. Each of these purposes may require (or prefer) different methodological approaches, and there is a need to determine what forms of replication can be performed for each. Furthermore, the underlying theories that are used to justify the interpretations of the results [29] should be further analyzed to identify the appropriate role for replication.

5.2 Moving Forward as a Field

Over the last decade, the Computing Education research field has seen a clear increase in research articles [45]. This shift was, in part, facilitated by conference program committees who sought more rigorous research over novel ideas, approaches and systems. For example, in the 2004 Koli Calling foreword, Lauri Malmi wrote:

> This year the program committee decided to call separately for research papers and discussion papers to make a clearer distinction between papers that present novel ideas, approaches and systems for CS education, and papers in which these issues have been elaborated further in some rigid research setting. Both types of papers are, however, equally necessary for the whole CS education community. New ideas and tools are the fuel for research work, and research is needed to convince us that we are really making progress towards our goal of improving learning. [28]

There is a need to take another step and to emphasize the need to replicate prior findings with the goal of generalizing and validating them. As a start, we propose that the Computing Education research community adopt a view similar to that of HCI, where guidelines for reviewers responsible for assessing submissions state[2]:

> Novelty is highly valued at CHI, but constructive replication can also be a significant contribution to human-computer interaction, and a new interpretation or evaluation of previously-published ideas can make a good CHI paper.

We argue that, if replication studies were explicitly solicited in CFPs and if their value were acknowledged in instructions to reviewers in our field, we would see more such papers. This would lead to more critical evaluation of new theories and better understanding of the generalizability of research findings.

5.3 Threats to Validity

While this study suggests that computer science education researchers hold similar beliefs about replication to researchers in other fields, potential issues with sampling and the internal validity of the survey instrument suggest that the results should be interpreted carefully and should not be used, in isolation, to make generalizations.

First, we do not have the ability to identify the biases in the sample. While the survey was distributed broadly, including the main list for our discipline's international professional organization, it is possible – and even likely – that respondents are both predisposed to being sensitive to replication work and have had experience with replication in the past. It is also likely, given the geographic distribution of respondents, that this survey predominantly reflects the state of the North American community, even though significant computer science education communities exist in Oceania and Europe. It is also difficult to determine what fraction of the total, active computer science education community is reflected in the response size (n=73) and whether the respondents accurately represent the community, as they self-identified themselves as active researchers.

Second, while the survey instrument was trialed before use, it was not formally validated, and respondents may have answered questions differently than intended. For example, respondents may have chosen to answer the questions in Part 2 of the survey as they believe the community *should be* as opposed to how the community *is*. Another issue is the use of the word "original" where "novel" may have been a better choice. Many respondents appeared to read "original" as "novel", as intended, but it's possible that some interpreted "original" as being a primary source work.

Finally, while the analysis of the free response sections of the survey have been reviewed by multiple authors, the interpretation of the results – and the selection of results presented – may reflect the biases of the group.

6. CONCLUSIONS

We have presented the results of a survey of 73 computer education researchers' perceptions of replication research. The survey responses suggest that researchers in our field believe that originality, defined as novelty – the documentation of previously undocumented phenomena, is a highly desirable property of research work. Original works are believed to have more impact, to be more prestigious, and to be more likely to be accepted for publication. These beliefs are similar to those of researchers in other fields that have suffered from crises caused by problems with replication in the field.

In response, we believe that, as a community, we must expand our definition of "originality." In contrast to the use of originality as foil to replication, "originality" required by journals is simply a requirement that the work being described be performed by the researchers writing the article. A paper about a strict replication is, by this definition, "original" if the authors completed the replication; the research question, which revolves around verification of a previously published phenomenon, is clearly stated; and the methodology, results, and implications are discussed. "Original" research advances knowledge of the field, even – and maybe even especially – if that knowledge is a confirmation or refutation of prior work.

We also believe that the community should adopt a standard that requires independent verification before an idea is accepted. Survey respondents agreed that verifiability is a desirable property of published work but doubt that many published projects in the computing education field are verifiable. Furthermore, they are, themselves, often unwilling to perform the work of verification themselves. Instead, because of the higher perceived value of "original" work, the discipline appears willing to build on individual studies.

While this issue is present in both qualitative and quantitative work, the privileged place accorded to statistically significant evidence is notable. These concerns are not new. In 1978, Carver argued that we should place more emphasis on the careful examination of data and replication of results than on statistical significance testing [8]. He noted that statistically significant results obtained by chance are unlikely to replicate, but robust phenomena will be replicable – with statistically significant results, if that is the evidence used – across many contexts.

In this paper as well as in the survey, we have argued that as a community, we use the terms replicate, reproduce and repeat interchangeably. To adopt a standard that requires replication, the computer science education community, like other scientific and social scientific disciplines, must define what replication means in our discipline. (One possible starting point could be R.A.P. taxonomy published in the ITiCSE working group report from 2015 by Ihantola et al [18].) Replication is potentially more difficult in social sciences, such as education, so generating a working definition of the various types of replication that can exist in our field and recognizing when to apply them is a necessary step to support a community-wide dialogue about the replication problem. Once accepted definitions are in place, we can begin to adopt practices that support and incentivize replication projects. What, specifically, those practices are is dependent on our context and will require effort to identify and adopt. But at the least, we should consider how to support venues for presenting and publishing replication work and should educate new members of our community about replication and the issues involved in completing an effective replication study.

[2] https://chi2016.acm.org/wp/guide-to-reviewing-papers-and-notes/

7. REFERENCES

[1] A. Ahadi, R. Lister, H. Haapala, and A. Vihavainen. Exploring machine learning methods to automatically identify students in need of assistance. In *Proceedings of the Eleventh Annual International Conference on International Computing Education Research*, ICER '15, pages 121–130, New York, NY, USA, 2015. ACM.

[2] A. Al-Zubidy, J. C. Carver, S. Heckman, and M. Sherriff. A (updated) review of empiricism at the sigcse technical symposium. In *Proceedings of the 47th ACM Technical Symposium on Computing Science Education*, SIGCSE '16, pages 120–125, New York, NY, USA, 2016. ACM.

[3] C. J. Anderson, Š. Bahník, M. Barnett-Cowan, F. A. Bosco, J. Chandler, C. R. Chartier, F. Cheung, C. D. Christopherson, A. Cordes, E. J. Cremata, et al. Response to comment on "estimating the reproducibility of psychological science". *Science*, 351(6277):1037–1037, 2016.

[4] J. B. Asendorpf, M. Conner, F. De Fruyt, J. De Houwer, J. J. Denissen, K. Fiedler, S. Fiedler, D. C. Funder, R. Kliegl, B. A. Nosek, et al. Recommendations for increasing replicability in psychology. *European Journal of Personality*, 27(2):108–119, 2013.

[5] M. Baker. Is there a reproducibility crisis? *Nature*, 533:452–454, May 2016.

[6] J. Barshay. Education researchers don't check for errors – dearth of replication studies. http://educationbythenumbers.org/content/education-researchers-dont-check-errors-dearth-replication-studies_1762/. Accessed: 2016-08-04.

[7] C. G. Begley and L. M. Ellis. Drug development: Raise standards for preclinical cancer research. *Nature*, 483(7391):531–533, 2012.

[8] R. Carver. The case against statistical significance testing. *Harvard Educational Review*, 48(3):378–399, 1978.

[9] T. Clear. Valuing computer science education research? In *Proceedings of the 6th Baltic Sea Conference on Computing Education Research: Koli Calling 2006*, Baltic Sea '06, pages 8–18, New York, NY, USA, 2006. ACM.

[10] O. S. Collaboration et al. Estimating the reproducibility of psychological science. *Science*, 349(6251):aac4716, 2015.

[11] O. J. Dunn. Multiple comparisons among means. *Journal of the American Statistical Association*, 56(293):52–64, 1961.

[12] H. Evanschitzky, C. Baumgarth, R. Hubbard, and J. S. Armstrong. Replication research's disturbing trend. *Journal of Business Research*, 60(4):411–415, 2007.

[13] D. Fanelli. Negative results are disappearing from most disciplines and countries. *Scientometrics*, 90(3):891–904, 2011.

[14] D. T. Gilbert, G. King, S. Pettigrew, and T. Wilson. Comment on "estimating the reproducibility of psychological science". *Science*, 351(6277):1037, 2016.

[15] M. A. Golden. Replication and non-quantitative research. *PS: Political Science & Politics*, 28(03):481–483, 1995.

[16] O. S. Gómez, N. Juristo, and S. Vegas. Replication, reproduction and re-analysis: Three ways for verifying experimental findings. In *Proceedings of the 1st international workshop on replication in empirical software engineering research (RESER 2010), Cape Town, South Africa*, 2010.

[17] M. Guzdial. Sigcse 2016 preview: Miranda parker replicated the fcs1. https://computinged.wordpress.com/2016/03/02/sigcse-2016-preview-miranda-parker-replicated-the-fcs1/. Accessed: 2016-08-05.

[18] P. Ihantola, A. Vihavainen, A. Ahadi, M. Butler, J. Börstler, S. H. Edwards, E. Isohanni, A. Korhonen, A. Petersen, K. Rivers, M. A. Rubio, J. Sheard, B. Skupas, J. Spacco, C. Szabo, and D. Toll. Educational data mining and learning analytics in programming: Literature review and case studies. In *Proceedings of the 2015 ITiCSE on Working Group Reports*, ITICSE-WGR '15, pages 41–63, New York, NY, USA, 2015. ACM.

[19] J. P. Ioannidis. Contradicted and initially stronger effects in highly cited clinical research. *Jama*, 294(2):218–228, 2005.

[20] J. P. Ioannidis. Why most published research findings are false. *PLoS Med*, 2(8):e124, 2005.

[21] J. P. Ioannidis, M. R. Munafo, P. Fusar-Poli, B. A. Nosek, and S. P. David. Publication and other reporting biases in cognitive sciences: detection, prevalence, and prevention. *Trends in Cognitive Sciences*, 18(5):235–241, 2014.

[22] J. Leinonen, K. Longi, A. Klami, and A. Vihavainen. Automatic inference of programming performance and experience from typing patterns. In *Proceedings of the 47th ACM Technical Symposium on Computing Science Education*, SIGCSE '16, pages 132–137, New York, NY, USA, 2016. ACM.

[23] R. Lister, T. Clear, Simon, D. J. Bouvier, P. Carter, A. Eckerdal, J. Jacková, M. Lopez, R. McCartney, P. Robbins, O. Seppälä, and E. Thompson. Naturally occurring data as research instrument: Analyzing examination responses to study the novice programmer. *SIGCSE Bull.*, 41(4):156–173, Jan. 2010.

[24] D. T. Lykken. Statistical significance in psychological research. *Psychological bulletin*, 70(3p1):151, 1968.

[25] A. Mackey. Why (or why not), when and how to replicate research. *Replication research in applied linguistics*, 2146, 2012.

[26] M. C. Makel and J. A. Plucker. Facts are more important than novelty replication in the education sciences. *Educational Researcher*, page 0013189X14545513, 2014.

[27] M. C. Makel, J. A. Plucker, and B. Hegarty. Replications in psychology research how often do they really occur? *Perspectives on Psychological Science*, 7(6):537–542, 2012.

[28] L. Malmi. Foreword to proceedings of the fourth finnish/baltic sea conference on computer science education, 2004.

[29] L. Malmi, J. Sheard, Simon, R. Bednarik, J. Helminen, P. Kinnunen, A. Korhonen, N. Myller, J. Sorva, and A. Taherkhani. Theoretical underpinnings of

computing education research: What is the evidence? In *Proceedings of the Tenth Annual Conference on International Computing Education Research*, ICER '14, pages 27–34, New York, NY, USA, 2014. ACM.

[30] L. Malmi, J. Sheard, Simon, R. Bednarik, J. Helminen, A. Korhonen, N. Myller, J. Sorva, and A. Taherkhani. Characterizing Research in Computing Education: A Preliminary Analysis of the Literature. In *Proceedings of the Sixth International Workshop on Computing Education Research*, ICER '10, pages 3–12, New York, NY, USA, 2010. ACM.

[31] J. N. Matias, S. Dasgupta, and B. M. Hill. Skill progression in scratch revisited. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, pages 1486–1490, New York, NY, USA, 2016. ACM.

[32] R. McCartney, J. Boustedt, A. Eckerdal, K. Sanders, and C. Zander. Can first-year students program yet?: A study revisited. In *Proceedings of the Ninth Annual International ACM Conference on International Computing Education Research*, ICER '13, pages 91–98, New York, NY, USA, 2013. ACM.

[33] R. McCauley, B. Hanks, S. Fitzgerald, and L. Murphy. Recursion vs. iteration: An empirical study of comprehension revisited. In *Proceedings of the 46th ACM Technical Symposium on Computer Science Education*, SIGCSE '15, pages 350–355, New York, NY, USA, 2015. ACM.

[34] R. Moonesinghe, M. J. Khoury, and A. C. J. Janssens. Most published research findings are false – but a little replication goes a long way. *PLoS Med*, 4(2):e28, 2007.

[35] J. R. Muma. The need for replication. *Journal of Speech, Language, and Hearing Research*, 36(5):927–930, 1993.

[36] M. C. Parker and M. Guzdial. Replicating a validated cs1 assessment (abstract only). In *Proceedings of the 47th ACM Technical Symposium on Computing Science Education*, SIGCSE '16, pages 695–695, New York, NY, USA, 2016. ACM.

[37] R. D. Peng. Reproducible research in computational science. *Science*, 334(6060):1226–1227, 2011.

[38] A. Petersen, J. Spacco, and A. Vihavainen. An Exploration of Error Quotient in Multiple Contexts. In *Proceedings of the 15th Koli Calling International Conference on Computing Education Research*, Koli Calling '15, New York, NY, USA, 2015. ACM.

[39] L. Porter, C. Bailey Lee, B. Simon, and D. Zingaro. Peer instruction: Do students really learn from peer discussion in computing? In *Proceedings of the Seventh International Workshop on Computing Education Research*, ICER '11, pages 45–52, New York, NY, USA, 2011. ACM.

[40] J. Randolph, G. Julnes, E. Sutinen, and S. Lehman. A methodological review of computer science education research. *Journal of Information Technology Education*, 7(1):135–162, 2008.

[41] J. J. Randolph and R. Bednarik. Publication bias in the computer science education research literature. *Journal of Universal Computer Science*, 14(4):575–589, 2008.

[42] S. Schmidt. Shall we really do it again? the powerful concept of replication is neglected in the social sciences. *Review of General Psychology*, 13(2):90, 2009.

[43] J. W. Schooler. Metascience could rescue the 'replication crisis'. *Nature*, 515(7525):9–9, 2014.

[44] O. Seppälä, P. Ihantola, E. Isohanni, J. Sorva, and A. Vihavainen. Do we know how difficult the rainfall problem is? In *Proceedings of the 15th Koli Calling Conference on Computing Education Research*, Koli Calling '15, pages 87–96, New York, NY, USA, 2015. ACM.

[45] Simon. *Emergence of computing education as a research discipline*. PhD thesis, 2015.

[46] B. A. Spellman. Introduction to the special section data, data, everywhere... especially in my file drawer. *Perspectives on Psychological Science*, 7(1):58–59, 2012.

[47] L. Thomas, A. Eckerdal, R. McCartney, J. E. Moström, K. Sanders, and C. Zander. Graduating students' designs: Through a phenomenographic lens. In *Proceedings of the Tenth Annual Conference on International Computing Education Research*, ICER '14, pages 91–98, New York, NY, USA, 2014. ACM.

[48] I. Utting, A. E. Tew, M. McCracken, L. Thomas, D. Bouvier, R. Frye, J. Paterson, M. Caspersen, Y. B.-D. Kolikant, J. Sorva, and T. Wilusz. A fresh look at novice programmers' performance and their teachers' expectations. In *Proceedings of the ITiCSE Working Group Reports Conference on Innovation and Technology in Computer Science Education-working Group Reports*, ITiCSE -WGR '13, pages 15–32, New York, NY, USA, 2013. ACM.

[49] D. W. Valentine. Cs educational research: A meta-analysis of sigcse technical symposium proceedings. *SIGCSE Bull.*, 36(1):255–259, Mar. 2004.

[50] C. Watson and F. W. Li. Failure rates in introductory programming revisited. In *Proceedings of the 2014 Conference on Innovation &#38; Technology in Computer Science Education*, ITiCSE '14, pages 39–44, New York, NY, USA, 2014. ACM.

APPENDIX

A. SURVEY

Respondents had the option of declining to answer any question.

Part 1: Demographics and Research / Teaching Focus

The questions 4, 5 and 7 had answer options *Yes*, *No* and *I Decline to answer this question.*

1. Where have you spent the majority of your career?
   - Africa
   - Asia
   - Australia/Oceania
   - Europe
   - North America
   - South America
   - Other
2. How many years of experience do you have as a teacher and/or researcher at the university level
   - 0-5
   - 5-10
   - 11-15
   - 15-20
   - 20-30
   - 30+
3. Are you a:
   - University teacher or researcher (professor, lecturer, etc.)
   - PhD student
   - Masters student / Sessional teaching staff
   - Other
4. Do you currently teach?
5. Do you currently do research?
6. What subjects do you teach most commonly?
7. Do you do computing education research?
8. What is your current main focus in research (not necessarily computing education)?
9. If you have experience conducting research in other fields, what are those fields?

Part 2: Replication and Computing Education Research

The following questions should be answered from the perspective of computing education research. For each question, the respondent could answer using five-level Likert scale options ranging from *strongly disagree* to *strongly agree*, or opt out from that specific question and choose *I decline to answer this question.*

1. Replication studies are harder to publish than original studies.
2. Replication studies are easier to conduct than original studies.
3. Original studies are more prestigious than replication studies.
4. Replication studies are too time-consuming for the recognition and reward.
5. Replication studies take energy and resources directly away from projects that reflect original thinking.
6. Original findings are generally harder to publish than replicated findings.
7. Replication studies are mechanical exercises, rather than major contributions to the field.
8. Replication studies bring less recognition and reward, including grant money, to their authors.
9. Original studies obtain more citations than replication studies.
10. To get tenure / to be promoted, a researcher needs a significant number of publications.
11. To get tenure / to be promoted, a researcher needs to publish more original than replication studies.
12. Replication studies do not count towards tenure or promotion.
13. To get tenure / to be promoted, a researcher needs to publish in peer reviewed venues.
14. To get tenure / to be promoted, a researcher's work must be highly cited by others in the field.
15. To get tenure / to be promoted, a researcher's work must be validated by others in the field.
16. Originality is an important criteria for evaluating work submitted for publication.
17. Reproducibility is an important criteria for evaluating work submitted for publication.
18. Verifiability is an important criteria for evaluating work submitted for publication.
19. Work submitted for publication should be grounded in previously published results.

Part 3: Replication Experiences

The questions 2–7 had answer options *Yes*, *No* and *I Decline to answer this question.* The respondent was instructed to answer either 8 or 9.

1. In your opinion, what proportion of published results in computing education research are reproducible? (i.e. the results of a given study could be replicated exactly or reproduced in multiple similar experimental systems with variations of experimental settings such as materials and experimental model)
   - 10%
   - 20%
   - 30%
   - 40%
   - 50%
   - 60%
   - 70%
   - 80%
   - 90%
   - 100%
2. Have you ever failed to reproduce one of your own results?
3. Have you ever failed to reproduce someone else's result?
4. Have you ever published a successful attempt to reproduce someone else's work?
5. Have you ever published a failed attempt to reproduce someone else's work?
6. Have you ever failed to publish a successful reproduction?
7. Have you ever failed to publish an unsuccessful reproduction?
8. Please explain why you choose to conduct replication studies.
9. Please explain why you choose not to conduct replication studies.

Part 4: Types of Replication Studies

In the above, we have used only the term replication to cover possible different kind of studies. Do you see any difference among studies that aim to replicate, reproduce or repeat some earlier research?