A Systematic Investigation of Replications in Computing Education Research

QIANG HAO, DAVID H. SMITH IV, NAITRA IRIUMI, and MICHAIL TSIKERDEKIS,
Western Washington University
ANDREW J. KO, University of Washington

As the societal demands for application and knowledge in computer science (CS) increase, CS student enrollment keeps growing rapidly around the world. By continuously improving the efficacy of computing education and providing guidelines for learning and teaching practice, computing education research plays a vital role in addressing both educational and societal challenges that emerge from the growth of CS students. Given the significant role of computing education research, it is important to ensure the reliability of studies in this field. The extent to which studies can be replicated in a field is one of the most important standards for reliability. Different fields have paid increasing attention to the replication rates of their studies, but the replication rate of computing education was never systematically studied. To fill this gap, this study investigated the replication rate of computing education between 2009 and 2018. We examined 2,269 published studies from three major conferences and two major journals in computing education, and found that the overall replication rate of computing education was 2.38%. This study demonstrated the need for more replication studies in computing education and discussed how to encourage replication studies through research initiatives and policy making.

CCS Concepts: • **Social and professional topics → Computing education**;

Additional Key Words and Phrases: Computing education, replications, computer science education, replication rate, assessment, evaluation, content analysis, educational policy, research methodology

ACM Reference format:
Qiang Hao, David H. Smith IV, Naitra Iriumi, Michail Tsikerdekis, and Andrew J. Ko. 2019. A Systematic Investigation of Replications in Computing Education Research. *ACM Trans. Comput. Educ.* 19, 4, Article 42 (August 2019), 18 pages.
https://doi.org/10.1145/3345328

1 BACKGROUND

In recent years, computer science (CS) has attracted rapidly growing numbers of students around the world. For instance, undergraduate CS enrollment has doubled from 2011 to 2017 in U.S. colleges [1]. The growth is not only limited to college students. In the United Kingdom and Australia, CS has been mandated for students in elementary and middle school [2, 3]. In the United States,

Authors' addresses: Q. Hao, CF 457, Western Washington University, 516 High Street, Bellingham, WA; email: qiang.hao@wwu.edu; D. H. Smith IV and N. Iriumi, CF 414, Western Washington University, 516 High Street, Bellingham, WA; emails: {smithd77, iriumin}@wwu.edu; M. Tsikerdekis, CF 489, Western Washington University, 516 High Street, Bellingham, WA; email: Michael.Tsikerdekis@wwu.edu; A. J. Ko, Mary Gates Hall 015E, University of Washington, Seattle, WA; email: ajko@uw.edu.

1946-6226/2019/08-ART42 $15.00
https://doi.org/10.1145/3345328

76% of high schools offer some form of CS learning opportunity, 60% offer at least one CS course, and 40% offer classes involving some programming [4]. Challenging questions of computing education emerge as the student population grows rapidly. These questions include, but are not limited to, how to increase student learning efficacy at scale, how to develop better tools to support learning and teaching of programming, and how to cultivate diversity and inclusion [5]. The prospect of answering such questions continues to attract an increasing number of computing education researchers and practitioners. Answering such questions requires not only novel ideas but also careful scrutiny of all ideas. Paraphrasing Carl Sagan, it is the ruthless skeptical scrutiny of all ideas that sets truth from nonsense [6].

The scrutiny of an idea requires replication of prior studies. Replication refers to purposeful repetition of prior studies to either confirm or disconfirm their findings [7]. If a hypothesis or theory is to be grounded in reality, the observations that the hypothesis or theory is based on must be replicable. Despite being known as the cornerstone of science, replications of prior studies are rarely attempted by researchers. Concerns regarding low replication rates received great attention in fields such as biology, medicine, economics, and psychology [8–12]. In the early 2010s, Begley and Ellis [8] attempted to fully replicate 53 highly cited cancer trial studies but only achieved a success rate of 11%. Similarly, Brian Nosek and 269 co-authors [9] attempted to replicate a total of 100 high-profile psychological publications but found that statistically significant effects could not be reproduced in 61 of the 100 studies. Such findings fully kicked off the replication crisis in science. Investigations of replications of different fields followed up, and most findings indicated that replication rates were extremely low. In addition, researchers in many fields rarely conducted replication studies for various reasons. For instance, Makel and Plucker [10] studied the replication rate of the top 100 educational journals ranked by 5-year impact factor but found that only 0.13% of published educational studies were replication studies. Surprisingly, 63.7% of all identified replication studies were conducted by the authors of the study being replicated. Moreover, 82% of the same-author replication studies successfully supported their predicted results. Hyperaccuracy like this is likely due to reasons such as inherent bias in research design, collecting data until the desired results emerge, and eliminating data that fail to support the hypothesis [10, 13].

Computing education shares similarities with other educational fields but also bears some significant differences. A noticeable similarity is the reliance on randomized-controlled experiment and quasi experimental design [14]. As in other educational fields, computing education researchers use randomized-controlled or quasi experiments for their research studies. As it is pointed out by Makel and Plucker [10], such experiments, including meta-analysis of randomized-controlled experiments, suffer from limitations such as post hoc hypothesizing with known results and publishing only positive results.

One difference from other fields is that computing education, as an interdisciplinary field, borrows heavily from different disciplines such as human-computer interaction (HCI) and software engineering [15]. HCI and software engineering have a clearer emphasis on replication than most educational fields [15]. For instance, the review guidelines of the Conference on Human Factors in Computing Systems (CHI) [16] state the following:

> "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."

As such, it is possible that the influences of HCI and software engineering lead to more replication studies in computing education than in other educational fields. However, evidence on this effect has yet to be established.

The most important difference between computing education and other educational fields lies in the publication channels. Only journals with high impact factors are considered premium publication channels by most educational fields [10]. In contrast, both journals and conference proceedings in computing education serve archival purposes, and both venues are deemed equally important.

Despite these differences between computing education and other fields, no systematic investigation on replication has been conducted so far for computing education. Ahadi et al. [17] surveyed 73 computing education researchers and found that they valued novelty the most. Although participants of the survey agreed that published research should be verified, they were reluctant to perform such research themselves. However, the extent to which this study's survey results can be generalized is still unknown. To fill this gap, this study investigates the replication rate of computing education research in the past decade. The results of this study contribute to a comprehensive understanding of replication in computing education research and emphasize the importance of publication guidelines that support replication studies.

2 DEFINITIONS OF REPLICATION

Replication refers to purposeful repetition of prior studies to either confirm or disconfirm their findings [7, 18]. The purposes of replication include the following:

- Controlling for fraud or artifacts
- Controlling for sample or statistical errors
- Assessing a hypothesis or theory from prior studies
- Testing the generalizability of a hypothesis or theory.

Lykken [19] first proposed three types of replications based on researcher interest in conducting empirical studies. The proposed three replication types include literal, operational, and constructive replications:

- A *literal* replication is an exact duplication of a prior study, including sampling techniques, experiment design, controls of the context, sample composition, measurement, and data analysis.
- An *operational* replication refers to an effort to duplicate the sampling techniques and experiment procedures of a prior study.
- A *constructive* replication refers to the effort to test prior findings using different experimental designs, measurements, and data analysis techniques that are more robust than prior studies.

It is worth noting that a literal replication of a prior study is almost impossible even for the same authors, because of the unavoidable differences between samples [10]. Schmidt [7] further simplified the duplication classification of Lykken [19] by eliminating the literal duplication and renaming the other two types as direct and conceptual replications:

- Corresponding to operational replication, *direct replication* refers to research efforts that stick to the same sampling techniques and experimental procedures.
- In contrast, *conceptual replications* correspond to constructive replication, which aims at both checking prior findings and overcoming limits of prior studies.

Schmidt's replication classification system was used more widely than that of Lykken [10]. To be consistent with studies investigating replication in other fields [e.g., 9, 10], we adopted the replication classification of Schmidt [7] in this study.

3 RESEARCH QUESTIONS

The research questions that guided this study include the following:

(1) What is the proportion of replication studies in computing education research publications?
(2) To what extent do replication studies in computing education successfully replicate prior findings?
(3) What is the research topic focus for replication studies in computing education?

4 METHOD

To answer the first research question, *What is the proportion of replication studies in computing education research publications?*, we collected papers from three major conference proceedings and two major journals in computing education between 2009 and 2018. The three computing education conferences included the following:

- Special Interest Group on Computer Science Education Technical Symposium (SIGCSE)
- International Computing Education Research Conference (ICER)
- Conference on Innovation and Technology in Computer Science Education (ITiCSE)

The two computing education journals included the following:

- ACM Transactions on Computing Education (TOCE)
- Computer Science Education Journal (CSEJ)

The included five publication venues are among the most selective and impactful in computing education research. The three conferences published 185 papers annually between 2009 and 2018.[1] The average acceptance rate of the three conferences between 2009 and 2018 was 34%. The calculation was performed by averaging 30 acceptance rates from the three conferences between 2009 and 2018.[2] The two journals published about 40 papers annually and tended to be even more selective than the three conferences. The number of articles from the five venues surpassed 2,200 between 2009 and 2018. As a result, we believe that the collected articles are representative of computing education research.

To estimate the replication rate, we used the search term $replicat[a-z]*$ to identify articles containing *replicate*, *replicating*, *replicated*, or *replication* at least one time. For each study, the term was searched against the whole article, including the title, keyword, abstract, and body. The true replication studies were further manually filtered from such articles. This method was used by many prior studies investigating replication rates [e.g., 10, 20, 21]. The manual filtering process was conducted by three raters, including two trained graduate students and one experienced computing education researcher. Two raters were tasked to conduct the first round of filtering. If there was a difference between the ratings on one study, a third rater would further discuss with the two raters and they would make a final decision collectively. The interrater reliability (Cohen's kappa) for this step was 0.96.

To answer the second research question, *To what extent do replication studies in computing education successfully replicate prior findings?*, each identified replication study was further analyzed in terms of

[1] The three conferences published 777 papers in 2018. Computing education as a field is growing rapidly.

[2] Different from other years, ITiCSE 2018 Working Group papers had a different acceptance rate from ITiCSE 2018 proceedings. As a result, the acceptance rate of ITiCSE 2018 Working Group papers was excluded from the calculation.

- whether it is a direct or conceptual replication,
- whether its findings are consistent with the replication target,
- whether it was conducted by the same authors as the replication target (having at least one overlapping author), and
- what methodology it adopts (i.e., quantitative, qualitative, or mixed method).

Similar to the manual filtering process described previously, the analysis was conducted by the three trained raters. Two raters were tasked to conduct the first round of analysis. If there was a difference between the ratings on one study, a third rater would further discuss with the two raters and they would make a final decision collectively. The average interrater reliability (Cohen's kappa) of the four analysis items was 0.88.

To answer the third research question, *What is the research topic focus for replication studies in computing education?*, the computing education research topic classifications proposed by Sheard et al. [22], Valentine [23], and Pears et al. [24] were synthesized, which yield a union of 18 topics in total. A few examples from the synthesized topic classification include Learning & Teaching Strategies, Assessment, and Learning Behavior. The synthesized topic classification was further applied to identify the topics of the replication studies. Like the manual filtering and replication analysis, the topic classification was conducted by the three trained raters. Two raters were tasked to conduct the first round of classification. If there was a difference between the ratings on one study, a third rater would further discuss with the two raters and they would make a final decision collectively. The interrater reliability (Cohen's kappa) was 0.86.

5 RESULTS

5.1 Replication in Computing Education Publications

To answer the first research question *What is the proportion of replication studies in computing education research publications?*, a total of 2,269 articles from SIGCSE, ITiCSE, ICER, TOCE, and CSEJ were collected as our dataset. The dataset was analyzed using the term $replicat[a - z]*$ (i.e., a regular expression describing variants of the word replication). A total of 370 articles in our dataset contained at least one occurrence of the term in either the title or the body of the article. Of these, 316 were identified as nonreplication studies through subsequent manual analysis. These studies had verbiage on replication, but most of them were simply stating the needs for replication of their own findings. Therefore, only 54 studies (2.38% of the dataset) were identified as true replication studies.

The overall replication rate of the five venues between 2009 and 2018 was 2.38% (54 articles). It is worth noting that the rate of replication studies has been increasing over time. For example, no articles in 2009 were identified as replication studies, whereas 3.1% of the articles published in 2018 were replication studies (Figure 1).

The average replication rate of computing education conferences (2.42%) is slightly higher than computing education journals (2.12%). The two journals (TOCE and CSEJ) published 330 studies between 2009 and 2018, of which 7 studies were identified as replication studies. In comparison, the three conferences (SIGCSE, ICER, and ITiCSE) published 1,939 studies between 2009 and 2018, of which 47 studies were identified as replication studies.

5.2 Replication Studies by Methodology, Replication Type, and Authorship

To answer the second research question, *To what extent do replication studies in computing education successfully replicate prior findings?*, we classified all identified replication studies by the adopted methodology (quantitative, mixed, and qualitative), replication type (direct vs. conceptual), and authorship (same vs. different).

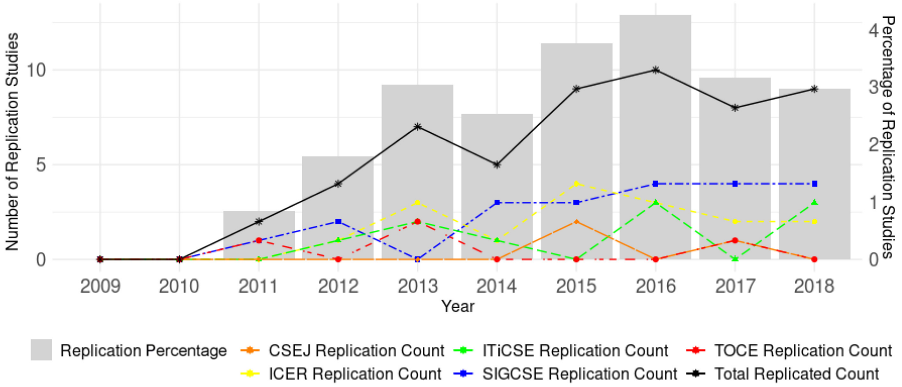

Fig. 1. Overview of replication studies from 2009 to 2018.

Table 1. Replication Studies by Methodology, Replication Type, and Authorship from 2009 to 2018

Methodology	Total	Success	Mixed	Failure
Quantitative	40	27 (68%)	9 (22%)	4 (10%)
Mixed	10	5 (50%)	5 (50%)	0 (0%)
Qualitative	4	2 (50%)	2 (50%)	0 (0%)
Replication Type	Total	Success	Mixed	Failure
Direct	13	7 (54%)	3 (23%)	3 (23%)
Conceptual	41	27 (66%)	3 (7%)	11 (27%)
Authorship	Total	Success	Mixed	Failure
Same authors	18	13 (72%)	1 (6%)	4 (22%)
Different authors	36	21 (58%)	5 (14%)	10 (28%)

Of all 54 replication studies, 24% were direct replications and 76% were conceptual replications. Thirty-four (63%) studies successfully replicated the original findings, whereas the rest reported failures or mixed results in their attempted replications.

When we grouped studies by the adopted methodologies (quantitative, mixed, and qualitative), it was evident that the majority of replication studies (74%) were quantitative. In contrast, 18.5% of replication studies used mixed methods, and only 7% were qualitative studies.

When we grouped studies by replication type (direct vs. conceptual), no clear patterns emerged in terms of the replication success rate. However, a clear pattern emerged when we grouped studies by authorship (same vs. different). When a replication study was conducted by the same authors, the replication success rate was 72% and the failure rate was 22%. In contrast, the success rate dropped to 58% and failure rate increased to 28% when all authors of a study were different from the authors of the study they were attempting to replicate (Table 1). This finding confirmed the same-author bias found in other fields [e.g., 9, 10].

5.3 Research Topics and Contexts in Replication Studies

To answer the third research question, *What is the research topic focus for replication studies in computing education?*, we classified all identified replication studies by both their topics and contexts. All identified replication studies were mapped to the synthesized computing education research topic classification. It is possible for one study to be mapped to more than one topic. For

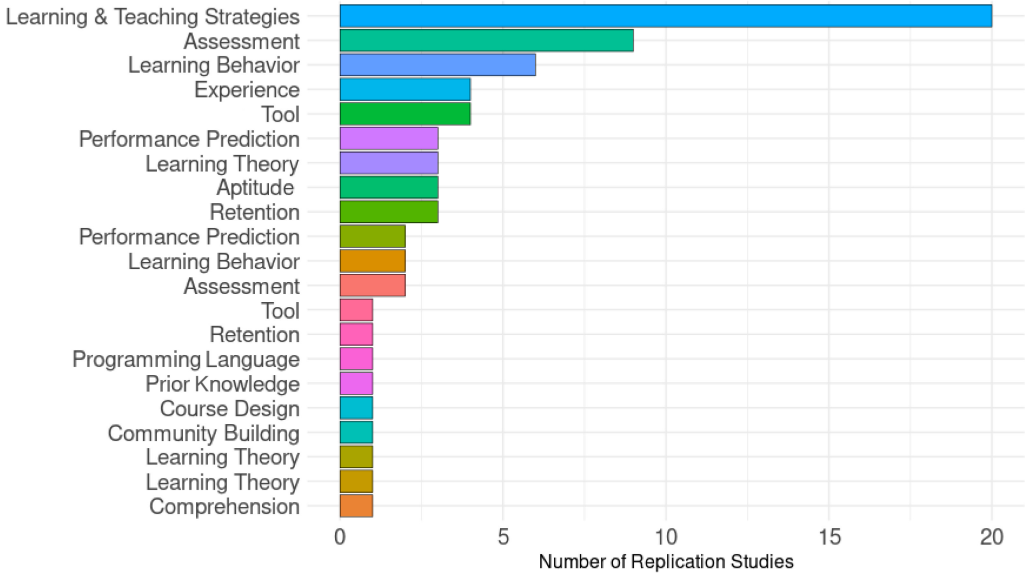

Fig. 2. Identified replication studies in computing education by topics.

instance, the study titled "Self-Efficacy, Cognitive Load, and Emotional Reactions in Collaborative Algorithms Labs—A Case Study" by Toma and Vahrenhold [25] was mapped into three topics, including (1) Learning & Teaching Strategies, (2) Assessment, and (3) Learning Theory. All covered topics are presented in Figure 2. Replication studies fell into the following top five topics:

- Learning & Teaching Strategies
- Assessment
- Learning Behavior
- Learning Theory
- Performance prediction

The identified top topics of replication studies aligned with the general trend of leading topics in computing education [26, 27].

Additionally, we also mapped all identified replication studies by research context. Five exclusive contexts were summarized, and each individual study fell into only one context:

- K-12: Computing education in K-12 contexts
- CS1: Computing education at the college level with exclusive focus on CS1
- Undergraduate: Computing education at the college level in general, with no focus on CS1
- Graduate: Computing education at the graduate level
- Others: Computing education in other contexts, such as in the workplace or among faculties.

The ratio comparison of the contexts is presented in Figure 3. It is worth noting that the majority of replication studies fall into the context of undergraduate computing education. Only six replication studies were in the context of K-12 computing education.

6 DISCUSSION

Our work sought to investigate the replication rates in computing education by surveying all publications from three major conferences and two major journals in this field between 2009 and

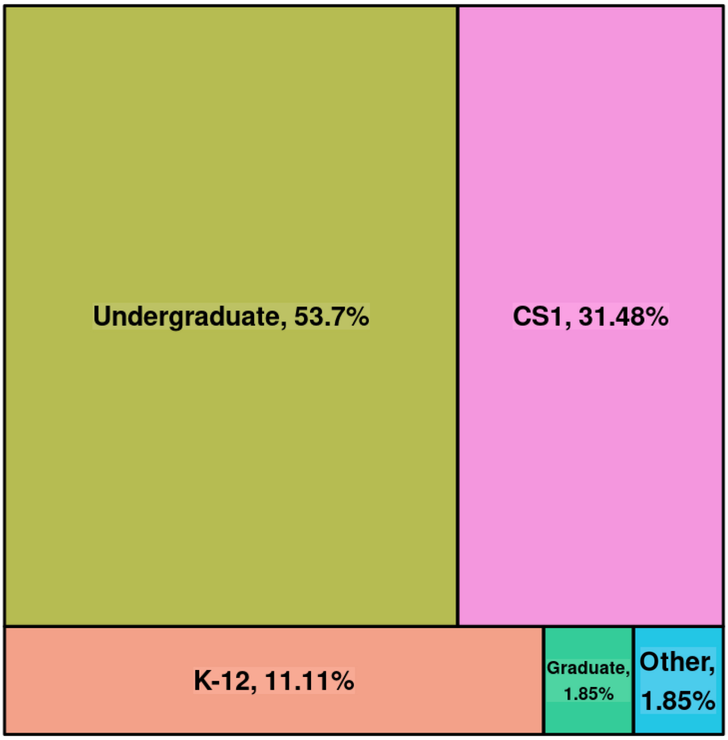

Fig. 3. Identified replication studies in computing education by context.

2018. We discovered that the overall replication rate was 2.38% (54 of 2,269 articles). Although a comprehensive replication rate cannot be found for many other scientific domains, this low replication rate of computing education is similar to psychology, business, and biology (1%–3%) [21, 28, 29].

However, it is worth noting that this rate is substantially higher than the reported overall replication rate of the top 100 journals ranked by impact factors (0.13%) in general education [10]. Even if only journals are considered, computing education has a noticeably higher replication rate than general education. The average replication rate of two major computing education journals (TOCE and CSEJ) between 2009 and 2018 was found to be 2.12%. This difference could be due to the fact that the replication rate of educational journals was reported before 2010 when less attention was paid to positive bias and scientific fraud. The difference could also be attributed to the practice of computing education researchers incorporating HCI and software engineering perspectives in their work, which in turn leads to adopting the emphasis on replication to computing education [e.g., 12, 30].

It is worth noting that the majority of identified replication studies were quantitative in nature. Only 4 studies in our dataset were qualitative, and only 10 studies used mixed methods. Considering that qualitative methods are widely used in computing education studies, this finding is surprising. Prior studies that attempt to classify computing education estimated that about 30% to 40% of the published studies adopted qualitative methods [26, 27, 31]. Several reasons may contribute to the extremely small number of qualitative studies being replicated in computing education, including (1) the lack of awareness among qualitative researcher in computing education in the importance of replication, (2) many qualitative computing education studies did not provide sufficient information for their studies to be replicated, and (3) the belief that qualitative studies only serve the goal of providing contextualized understanding for unique study cases [32, 33].

Our findings further confirmed the findings in psychology and general education on same-author bias [e.g., 10, 21]. When there was no overlap in authorship, the replication success rate dropped substantially (from 72% to 58%). On one hand, the same-author replications might benefit from the experience and knowledge of conducting the same experiment once before, which contributes to further success. On the other hand, the difference may indicate the existence of potential bias or influences of uncontrollable environmental factors. Therefore, the repetition of similar findings from same authors in computing education research may deserve further investigation.

Additionally, we would like to highlight two key findings on the replication study topics. First, despite retention being studied intensively in the past decade, only four replication studies were found focusing on this topic. This finding is not about doubting the documented retention efforts. However, replication on this topic is critical to understanding what retention strategies or interventions can be generalized. This understanding is important for widespread adoption of effective strategies. Without replication, it is difficult to achieve the goal. Second, only six studies in our dataset were identified as replication studies conducted in the context of K-12 computing education. The expansion of computing education into K-12 has gained tremendous momentum in the past decade, and research in this context also gained increasing attention. However, compared to undergraduate computing education research, more replications are needed to confirm the generalizability of findings in the K-12 context.

The following sections will discuss the limitations of this study, challenges and opportunities in conducting replication studies in computing education, and policies that encourage more replication studies.

6.1 Limitations

This study is not without limitations. If an article is not framed as a replication study (i.e., using words such as *replicate*, *replicating*, *replicated*, or *replication*), we did not count it as a replication study. Some authors might conduct a replication study but not explicitly specify the intention. Such articles were excluded from the counting using the research method of this study. Although we made an effort to search for replication studies that avoided using the terms, we did not identify any. Without explicit clarification of the intention, a replication study may greatly limit readers of interest to make connections between studies answering the same (or similar) questions. Future research may consider an in-depth analysis of a smaller amount of articles on trending topics, which can give a more accurate estimate of the percentage of the implicit replication studies.

The counting of paper numbers may not accurately reflect the quantity of actual published research articles in the included conference proceedings (SIGCSE, ITiCSE, and ICER). The three major computing conference proceedings include various types of scholarly products, such as full papers, panels, student research competition, and doctoral consortium. To filter nonresearch products, only full papers were counted and examined from the conference proceedings. However, some of the full papers were experience reports, which describe computing education interventions and provide reflections on their efficacy [34]. Experience reports are not considered research papers, but there is no easy way to tell if a full paper is a research article or an experience report. As result, the total counts of research articles might be inflated, which may further lead to an underestimation of the replication rate. If computing education conferences can make a clear difference between full research articles and experience reports in the future archival process, it will help to better gauge the replication rate of the field. The distinction between research articles and experience reports may also help researchers new to this field to more easily navigate the literature.

Additionally, computing education conferences with shorter history or smaller scales (e.g., Koli Calling and Australasian Computing Education Conference) were not included in the counting of replication studies, given that they published significantly fewer computing education studies. The exclusion of such venues may limit the representativeness of this study. Future studies may consider an evaluation of the quality and replication rates of such venues.

6.2 Challenges of Conducting Replication Studies

Not all computing education researchers support replication studies. Many researchers have expressed hesitation in conducting replication studies because they believed that (1) the value of original works is more significant than replication studies or (2) original works carry more weight in determining promotion [17]. These beliefs are associated with the observation that few replication studies have been published and the expectation that researchers will focus on novelty. Although replication is not a panacea to all research problems, dismissing replications reflects a deep misunderstanding of science and a bias toward novelty over truth [10, 35]. Science is self-correcting at an extremely slow pace. Without deliberately replicating and checking important works, the self-correction may be random and even slower. If computing education research is to be used for establishing robust policy and practice guidelines, deliberate replication and winnowing are essential to guarantee the reliability and stability of research findings.

Some computing education researchers have expressed doubts about the necessity of replications due to the highly contextualized nature of their research [17]. Although computing education research bears significant differences from general education research, the research does share many similarities. The nature of high contextualization is a feature shared by all educational studies. Factors specific to student demographics, classroom climate, and institutional culture definitely play roles in any empirical educational studies [35]. As a result, it is important to verify if the detected effects are mainly due to such contextualized factors through replications. If a replication failure happens, and the failure is mainly due to uncontrollable environmental factors, the fragility of the original findings needs to be recorded and discussed. In other words, replications can help computing education as a field avoid the risk of disseminating overgeneralized or oversimplified findings.

6.3 Policies That Encourage Replications

The need to increase replication studies is urgent given the low replication rate of computing education research. Different solutions have been proposed in other fields, such as biology, medicine, and psychology. Among the various proposed solutions, we would like to highlight four of them that have the potential of benefiting computing education:

- Editorial and review policies could be revised to explicitly encourage replication studies.
- A crowd-sourcing project that tracks replication studies is needed.
- A healthy environment of replication studies needs to be cultivated and maintained.
- An ongoing discussion on what merits replication is needed.

First, editorial and review policies could be revised to explicitly encourage replication studies. Multiple major journals and conferences in other fields have initiated this change [36]. Conferences, as an important publication venue of computing education, may consider similar strategies, such as reserving a dedicated section for replication studies or setting explicit review guidelines to encourage submissions of replication studies. Reviewers may lack the knowledge of replications or not be aware of their value, especially direct replications. More elaborate review guidelines may help educate reviewers about the value of replication studies and see the difference between direct and conceptual replication studies. The Association for Computing Machinery (ACM) recently published Artifact Review and Badging, which described a review process and badge system that emphasizes replicable and reproducible research [37]. If computing education conferences or journals can adopt this review process, it is likely to promote the integrity of the whole computing education research ecosystem. It is worth noting that despite the emphasis on replicable and reproducible research, ACM Artifact Review and Badging does not attempt to give proper credits to replication studies or directly address the difficulty in publishing replication studies. To enhance replicability and reproduciblity of a research field, the encouragement of replication studies is inevitable. As an example, ICER has explicitly encouraged replication studies in its *Call for Participation* [38]. If other publication venues in computing education can follow this direction, it will help change the culture that emphasizes originality over credibility.

Second, a crowd-sourcing project that tracks replication studies is needed. If a small group of interested computing education researchers can contribute to a Wiki-based service dedicated to recording replication studies, it will help more researchers easily see what effects can be reproduced and provide effective guidance for practitioners who want to improve the efficacy of student teaching and learning. The U.S. Department of Education's Institute for Education Science has an initiative called *What Works Clearinghouse* (WWC), which selects studies by standards such as effect size, sample size, and whether it is a randomized-controlled experiment [14]. The initiative helps to filter out a subset of high-quality studies. However, even a randomized-controlled study with significantly large effect size may still suffer from various limitations [39]. In addition, most studies do not have access to meet the sample size criteria of WWC, but that does not mean that such studies were poorly conducted. Although it is arguably important that experiments have high statistical power to detect nontrivial effects, statistical power should not determine outcomes on its own. For example, many studies tend to reject the likelihood of an effect due to the lack of a large sample (i.e., a false negative—Type II error), whereas many other studies tend to confirm negligible effects simply because they are statistically significant (i.e., a false positive—Type I error). In both of those examples, study contexts should matter more than just results alone. Properly conducted studies with nonsignificant results should still be considered as valuable. Therefore, the efforts of WWC will be sufficiently complemented if domain-based educational research (e.g., computing education research) can initiate crowd-sourcing projects that track replication efforts.

Third, a healthy environment of replication studies needs to be cultivated and maintained. A replication, despite of its results, should "neither cement or condemn the original findings" [10]. An individual study by itself may lack the stability to fully verify or reject the original findings. Many replications conducted in a timely manner can better inform the studied research question. Most importantly, a replication failure should not be interpreted as a signal of fraud or research misconduct. Various reasons can contribute to unreplicable findings of a study, such as experiment design flaws, uncontrollable environmental factors, and mistakes in data analysis. That being said, computing education, moving forward as a field, needs to constantly winnow the true from the false, and weed out narrow findings. An environment that allows mistakes and being wrong is essential for replications.

Finally, an ongoing discussion on what merits replication is needed. Similar discussions have been ongoing in other fields, such as psychology and general education. Different fields tend to emphasize different aspects of replications. From an intellectual aspect, Hunt [40] recommended that studies that lack robust methodology or bear design flaws deserve replications. From a practical aspect, Makel et al. [21] recommended optimizing resources devoted to research and focusing on studies that have great potential in impacting policies and practice. One of the selection criteria could be the number of citations. More field-specific aspects should be considered other than intellectual and practical aspects, such as the most important research questions and critical needs of a field. Five areas have been identified as the most important in computing education, including broadening participation, computing in K-12, computing in STEM education, students and learning issues, and tools [5]. As this study found, replication studies in some areas, such as computing in K-12 and computing in STEM education, were very rare in the past decade. The importance of the research topic and the rarity of replications both contribute to the necessity of replications in such areas. Furthermore, computing education, as a younger field, faces challenges such as increasing research student numbers and effectively training students. Replication may serve as a good entry point for students to gain hands-on research experience. In that case, we could satisfy the needs for replication studies and benefit students at the same time.

7 CONCLUSION

Replication is a core principle of the scientific method. This study reveals that computing education, like many other fields, suffers from an overall low replication rate through systematic investigation. The 2.38% overall replication rate of computing education research is higher than journals in general education (0.13%) but similar to psychology, business, and biology (1%–3%).

When we analyzed the identified replication studies by venues and authorship, we confirmed some findings from other fields but also reached some findings unique to computing education research. The same-author bias found in other fields was confirmed in computing education research. When a replication study is conducted by the same author(s) rather than different authors, the success rate goes up from 58% to 72%. Although computation education journals in general tend to have a higher requirement for accepted papers than conferences, that does not render a noticeable difference on the replication rate between computing education journals and conferences. The replication rate of computing education conferences (2.42%), although slightly higher, is almost the same as computing education journals (2.12%).

When we analyzed the identified replication studies by adopted methodology and context, we reached some findings that are worth noting by all computing education researchers. Qualitative methods are common in published computing education research, but only four identified replication studies were qualitative in our dataset. Similarly, computing in K-12, as a key area of computing education, had almost no replication studies being identified in our dataset. More replication studies are essential to move the field forward, especially where replication is extremely lacking, such as studies adopting qualitative methods and studies in computing in K-12. If computing education research is to be relied on for establishing policies and practice guidelines, we need both initiatives that encourage replication studies and the involvement of more computing education researchers.

APPENDIX

The summary of identified replication studies in computing education can be accessed at https://github.com/Neo-Hao/replications-in-computing-education. The information of the summary includes the following:

- Publication information
- Studied topics
- Study contexts
- Shared authorship with replication targets
- Replication types
- Results compared to replication targets

Here is the list of the identified replication studies:

(1) Allison Elliott Tew, Brian Dorn, and Oliver Schneider. 2012. Toward a validated computing attitudes survey. In *Proceedings of the 2012 Annual ACM International Computing Education Research Conference*. ACM, New York, NY, 135–142.

(2) Robert McCartney, Jonas Boustedt, Anna Eckerdal, Kate Sanders, and Carol Zander. 2013. Can first-year students program yet? A study revisited. In *Proceedings of the 2012 Annual ACM International Computing Education Research Conference*. ACM, New York, NY, 91–98.

(3) Colleen M. Lewis, Huda Khayrallah, and Amy Tsai. 2013. Mining data from the AP CS A exam: Patterns, non-patterns, and replication failure. In *Proceedings of the 2012 Annual ACM International Computing Education Research Conference*. ACM, New York, NY, 115–122.

(4) Elizabeth Patitsas, Michelle Craig, and Steve Easterbrook. 2013. Comparing and contrasting different algorithms leads to increased student learning. In *Proceedings of the 2013 Annual ACM International Computing Education Research Conference*. ACM, New York, NY, 145–152.

(5) Briana B. Morrison, Brian Dorn, and Mark Guzdial. 2014. Measuring cognitive load in introductory CS: Adaptation of an instrument. In *Proceedings of the 2014 Annual ACM International Computing Education Research Conference*. ACM, New York, NY, 131–138.

(6) Matthew C. Jadud and Brian Dorn. 2015. Aggregate compilation behavior: Findings and implications from 27,698 users. In *Proceedings of the 2015 Annual ACM International Computing Education Research Conference*. ACM, New York, NY, 131–139.

(7) Adam S. Carter, Christopher D. Hundhausen, and Olusola Adesope. 2015. The normalized programming state model: Predicting student performance in computing courses based on programming behavior. In *Proceedings of the 2015 Annual ACM International Computing Education Research Conference*. ACM, New York, NY, 141–150.

(8) Stephen Cooper, Karen Wang, Maya Israni, and Sheryl Sorby. 2015. Spatial skills training in introductory computing. In *Proceedings of the 2015 Annual ACM International Computing Education Research Conference*. ACM, New York, NY, 13–20.

(9) Briana B. Morrison, Lauren E. Margulieux, and Mark Guzdial. 2015. Subgoals, context, and worked examples in learning computing problem solving. In *Proceedings of the 2015 Annual ACM International Computing Education Research Conference*. ACM, New York, NY, 21–29.

(10) Michael Hewner and Shitanshu Mishra. 2016. When everyone knows CS is the best major: Decisions about CS in an Indian context. In *Proceedings of the 2016 Annual ACM International Computing Education Research Conference*. ACM, New York, NY, 3–11.

(11) Miranda C. Parker, Mark Guzdial, and Shelly Engleman. 2016. Replication, validation, and use of a language independent CS1 knowledge assessment. In *Proceedings of the 2016 Annual ACM International Computing Education Research Conference*. ACM, New York, NY, 93–101.

(12) Briana B. Morrison, Adrienne Decker, and Lauren E. Margulieux. 2016. Learning loops: A replication study illuminates impact of HS courses. In *Proceedings of the 2016 Annual ACM International Computing Education Research Conference*. ACM, New York, NY, 221–230.

(13) Kathryn Cunningham, Sarah Blanchard, Barbara Ericson, and Mark Guzdial. 2017. Using tracing and sketching to solve programming problems: Replicating and extending an analysis of what students draw. In *Proceedings of the 2017 Annual ACM International Computing Education Research Conference*. ACM, New York, NY, 164–172.

(14) Briana B. Morrison. 2017. Dual modality code explanations for novices: Unexpected results. In *Proceedings of the 2017 Annual ACM International Computing Education Research Conference*. ACM, New York, NY, 226–235.

(15) Laura Toma and Jan Vahrenhold. 2018. Self-efficacy, cognitive load, and emotional reactions in collaborative algorithms labs—A case study. In *Proceedings of the 2018 Annual ACM International Computing Education Research Conference*. ACM, New York, NY, 1–10.

(16) Eva Marinus, Zoe Powell, Rosalind Thornton, Genevieve McArthur, and Stephen Crain. 2018. Unravelling the cognition of coding in 3-to-6-year olds: The development of an assessment tool and the relation between coding ability and cognitive compiling of syntax in natural language. In *Proceedings of the 2018 Annual ACM International Computing Education Research Conference*. ACM, New York, NY, 133–141.

(17) Paul Denny, Andrew Luxton-Reilly, and Ewan Tempero. 2012. All syntax errors are not equal. In *Proceedings of the 2012 Annual Conference on Innovation and Technology in Computer Science Education*. ACM, New York, NY, 75–80.

(18) Ville Isomöttönen, Ville Tirronen, and Michael Cochez. 2013. Issues with a course that emphasizes self-direction. In *Proceedings of the 2013 Annual Conference on Innovation and Technology in Computer Science Education*. ACM, New York, NY, 111–116.

(19) Jaime Spacco, Davide Fossati, John Stamper, and Kelly Rivers. 2013. Towards improving programming habits to create better computer science course outcomes. In *Proceedings of the 2013 Annual Conference on Innovation and Technology in Computer Science Education*. ACM, New York, NY, 243–248.

(20) Christopher Watson and Frederick W. B. Li. 2014. Failure rates in introductory programming revisited. In *Proceedings of the 2014 Annual Conference on Innovation and Technology in Computer Science Education*. ACM, New York, NY, 39–44.

(21) Renate Thies and Jan Vahrenhold. 2016. Back to school: Computer science unplugged in the wild. In *Proceedings of the 2016 Annual Conference on Innovation and Technology in Computer Science Education*. ACM, New York, NY, 118–123.

(22) Juho Leinonen, Krista Longi, Arto Klami, Alireza Ahadi, and Arto Vihavainen. 2016. Typing patterns and authentication in practical programming exams. In *Proceedings of the 2016 Annual Conference on Innovation and Technology in Computer Science Education*. ACM, New York, NY, 160–165.

(23) Jennifer Campbell, Diane Horton, and Michelle Craig. 2016. Factors for success in online CS1. In *Proceedings of the 2016 Annual Conference on Innovation and Technology in Computer Science Education*. ACM, New York, NY, 320–325.

(24) Keith Quille and Susan Bergin. 2018. Programming: Predicting student success early in CS1. A re-validation and replication study. In *Proceedings of the 2018 Annual Conference on Innovation and Technology in Computer Science Education*. ACM, New York, NY, 15–20.

(25) Soohyun Nam Liao, William G. Griswold, and Leo Porter. 2018. Classroom experience report on jigsaw learning. In *Proceedings of the 2018 Annual Conference on Innovation and Technology in Computer Science Education*. ACM, New York, NY, 302–307.

(26) Robert McCartney and Kate Sanders. 2018. ITiCSE working groups and collaboration in the computing education community. In *Proceedings of the 2018 Annual Conference on Innovation and Technology in Computer Science Education*. ACM, New York, NY, 332–337.

(27) Alex D. Radermacher and Gursimran S. Walia. 2011. Investigating the effective implementation of pair programming: An empirical investigation. In *Proceedings of the 2011 ACM SIGCSE Technical Symposium on Computer Science Education*. ACM, New York, NY, 655–660.

(28) Alan C. Jamieson, Lindsay H. Jamieson, and Angela C. Johnson. 2012. Application of non-programming focused Treisman-style workshops in introductory computer science. In *Proceedings of the 2012 ACM SIGCSE Technical Symposium on Computer Science Education.* ACM, New York, NY, 271–276.

(29) Laurie Murphy, Renee McCauley, and Sue Fitzgerald. 2012. Explain in plain English questions: Implications for teaching. In *Proceedings of the 2012 ACM SIGCSE Technical Symposium on Computer Science Education.* ACM, New York, NY, 385–390.

(30) Kuba Karpierz and Steven A. Wolfman. 2014. Misconceptions and concept inventory questions for binary search trees and hash tables. In *Proceedings of the 2014 ACM SIGCSE Technical Symposium on Computer Science Education.* ACM, New York, NY, 109–114.

(31) Neil Christopher Charles Brown, Michael Kölling, Davin McCall, and Ian Utting. 2014. Blackbox: A large scale repository of novice programmers' activity. In *Proceedings of the 2014 ACM SIGCSE Technical Symposium on Computer Science Education.* ACM, New York, NY, 223–228.

(32) Mark Zarb, Janet Hughes, and John Richards. 2014. Evaluating industry-inspired pair programming communication guidelines with undergraduate students. In *Proceedings of the 2014 ACM SIGCSE Technical Symposium on Computer Science Education.* ACM, New York, NY, 361–366.

(33) Paul Denny. 2015. Generating practice questions as a preparation strategy for introductory programming exams. In *Proceedings of the 2015 ACM SIGCSE Technical Symposium on Computer Science Education.* ACM, New York, NY, 278–283.

(34) Arto Vihavainen, Craig S. Miller, and Amber Settle. 2015. Benefits of self-explanation in introductory programming. In *Proceedings of the 2015 ACM SIGCSE Technical Symposium on Computer Science Education.* ACM, New York, NY, 284–289.

(35) Renee McCauley, Brian Hanks, Sue Fitzgerald, and Laurie Murphy. 2015. Recursion vs. iteration: An empirical study of comprehension revisited. In *Proceedings of the 2015 ACM SIGCSE Technical Symposium on Computer Science Education.* ACM, New York, NY, 350–355.

(36) Briana B. Morrison, Lauren E. Margulieux, Barbara Ericson, and Mark Guzdial. 2016. Subgoals help students solve Parsons problems. In *Proceedings of the 2016 ACM SIGCSE Technical Symposium on Computer Science Education.* ACM, New York, NY, 42–47.

(37) Juho Leinonen, Krista Longi, Arto Klami, and Arto Vihavainen. 2016. Automatic inference of programming performance and experience from typing patterns. In *Proceedings of the 2016 ACM SIGCSE Technical Symposium on Computer Science Education.* ACM, New York, NY, 132–137.

(38) Robert Deloatch, Brian P. Bailey, and Alex Kirlik. 2016. Measuring effects of modality on perceived test anxiety for computer programming exams. In *Proceedings of the 2016 ACM SIGCSE Technical Symposium on Computer Science Education.* ACM, New York, NY, 291–296.

(39) Daniel Zingaro and Leo Porter. 2016. Impact of student achievement goals on CS1 outcomes. In *Proceedings of the 2016 ACM SIGCSE Technical Symposium on Computer Science Education.* New York, NY, 279–296.

(40) David Weintrop and Nathan Holbert. 2017. From blocks to text and back: Programming patterns in a dual-modality environment. In *Proceedings of the 2017 ACM SIGCSE Technical Symposium on Computer Science Education.* ACM, New York, NY, 633—638.

(41) Yingjun Cao and Leo Porter. 2017. Evaluating student learning from collaborative group tests in introductory computing. In *Proceedings of the 2017 ACM SIGCSE Technical Symposium on Computer Science Education.* ACM, New York, NY, 99–104.

(42) Christine Alvarado, Mia Minnes, and Leo Porter. 2017. Micro-classes: A structure for improving student experience in large classes. In *Proceedings of the 2017 ACM SIGCSE Technical Symposium on Computer Science Education*. ACM, New York, NY, 21–26.

(43) Breanne K. Litts, Yasmin B. Kafai, Debora Lui, Justice Walker, and Sari Widman. 2017. Understanding high school students' reading, remixing, and writing codeable circuits for electronic textiles. In *Proceedings of the 2017 ACM SIGCSE Technical Symposium on Computer Science Education*. ACM, New York, NY, 381–386.

(44) Daniel Zingaro, Michelle Craig, Leo Porter, Brett A. Becker, Yingjun Cao, Phill Conrad, Diana Cukierman, Arto Hellas, Dastyni Loksa, and Neena Thota. 2016. Achievement goals in CS1: Replication and extension. In *Proceedings of the 2018 ACM SIGCSE Technical Symposium on Computer Science Education*. ACM, New York, NY, 687–692.

(45) Chris Wilcox and Albert Lionelle. 2018. Quantifying the benefits of prior programming experience in an introductory computer science course. In *Proceedings of the 2018 ACM SIGCSE Technical Symposium on Computer Science Education*. ACM, New York, NY, 80–85.

(46) Onni Aarne, Petrus Peltola, Juho Leinonen, and Arto Hellas. 2018. A study of pair programming enjoyment and attendance using study motivation and strategy metrics. In *Proceedings of the 2018 ACM SIGCSE Technical Symposium on Computer Science Education*. ACM, New York, NY, 759–764.

(47) Austin Cory Bart, Dennis Kafura, Clifford A. Shaffer, and Eli Tilevich. 2018. Reconciling the promise and pragmatics of enhancing computing pedagogy with data science. In *Proceedings of the 2018 ACM SIGCSE Technical Symposium on Computer Science Education*. ACM, New York, NY, 1029–1034.

(48) Brian Dorn and Allison Elliott Tew. 2015. Empirical validation and application of the computing attitudes survey. *Computer Science Education* 25, 1, 1–36.

(49) Jamie Payton, Tiffany Barnes, Kim Buch, Audrey Rorrer, and Huifang Zuo. 2015. The effects of integrating service learning into computer science: An inter-institutional longitudinal study. *Computer Science Education* 25, 3, 311–324.

(50) Jane G. Stout and Jennifer M. Blaney. 2017. "But it doesn't come naturally": How effort expenditure shapes the benefit of growth mindset on women's sense of intellectual belonging in computing. *Computer Science Education* 27, 3–4, 215–228.

(51) Uolevi Nikula, Orlena Gotel, and Jussi Kasurinen. 2011. A motivation guided holistic rehabilitation of the first programming course. *ACM Transactions on Computing Education* 11, 4, 24.

(52) Andreas Stefik and Susanna Siebert. 2013. An empirical investigation into programming language syntax. *ACM Transactions on Computing Education* 13, 4, 19.

(53) Cynthia Bailey Lee, Saturnino Garcia, and Leo Porter. 2013. Can peer instruction be effective in upper-division computer science courses? *ACM Transactions on Computing Education* 13, 3, 12.

(54) Adam S. Carter, Christopher D. Hundhausen, and Olusola Adesope. 2017. Blending measures of programming and social behavior into predictive models of student achievement in early computing courses. *ACM Transactions on Computing Education* 17, 3, 12.

REFERENCES

[1] Computing Research Association. 2017. Generation CS: Computer Science Undergraduate Enrollments Surge Since 2006. Retrieved July 24, 2019 from https://cra.org/data/generation-cs/.

[2] Neil C. C. Brown, Sue Sentance, Tom Crick, and Simon Humphreys. 2014. Restart: The resurgence of computer science in UK schools. *ACM Transactions on Computing Education* 14, 2 (2014), 9.

[3] Peter Hubwieser, Michail N. Giannakos, Marc Berges, Torsten Brinda, Ira Diethelm, Johannes Magenheim, Yogendra Pal, Jana Jackova, and Egle Jasute. 2015. A global snapshot of computer science education in K-12 schools. In *Proceedings of the 2015 ITiCSE on Working Group Reports.* ACM, New York, NY, 65–83.

[4] Google Inc. and Gallup Inc. 2016. Trends in the State of Computer Science in U.S. K-12 Schools. Retrieved July 24, 2019 from https://services.google.com/fh/files/misc/trends-in-the-state-of-computer-science-report.pdf.

[5] Steve Cooper, Shuchi Grover, Mark Guzdial, and Beth Simon. 2014. A future for computing education research. *Communications of the ACM* 57, 11 (2014), 34–36.

[6] Carl Sagan. 1999. *The Demon-Haunted World: Science as a Candle in the Dark*. Ballantine Books, New York, NY.

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

[8] C. Glenn Begley and Lee M. Ellis. 2012. Drug development: Raise standards for preclinical cancer research. *Nature* 483, 7391 (2012), 531.

[9] Open Science Collaboration. 2015. Estimating the reproducibility of psychological science. *Science* 349, 6251 (2015), aac4716.

[10] Matthew C. Makel and Jonathan A. Plucker. 2014. Facts are more important than novelty: Replication in the education sciences. *Educational Researcher* 43, 6 (2014), 304–316.

[11] Andreas Stefik and Stefan Hanenberg. 2014. The programming language wars: Questions and responsibilities for the programming language community. In *Proceedings of the 2014 ACM International Symposium on New Ideas, New Paradigms, and Reflections on Programming and Software.* ACM, New York, NY, 283–299.

[12] Neil C. C. Brown, Amjad Altadmri, Sue Sentance, and Michael Kölling. 2018. Blackbox, five years on: An evaluation of a large-scale programming data collection project. In *Proceedings of the 2018 ACM Conference on International Computing Education Research.* ACM, New York, NY, 196–204.

[13] Ed Yong. 2012. Bad copy. *Nature* 485, 7398 (2012), 298.

[14] What Works Clearinghouse. 2014. *Procedures and Standards Handbook (Version 3.0).* U.S. Department of Education: Washington, DC.

[15] Forrest J. Shull, Jeffrey C. Carver, Sira Vegas, and Natalia Juristo. 2008. The role of replications in empirical software engineering. *Empirical Software Engineering* 13, 2 (2008), 211–218.

[16] Association for Computing Machinery. 2019. Guide to Reviewing Papers. Retrieved July 24, 2019 from https://chi2019.acm.org/guide-to-reviewing-papers/.

[17] Alireza Ahadi, Arto Hellas, Petri Ihantola, Ari Korhonen, and Andrew Petersen. 2016. Replication in computing education research: Researcher attitudes and experiences. In *Proceedings of the 16th Koli Calling International Conference on Computing Education Research.* ACM, New York, NY, 2–11.

[18] John P. A. Ioannidis. 2005. Contradicted and initially stronger effects in highly cited clinical research. *JAMA* 294, 2 (2005), 218–228.

[19] David T. Lykken. 1968. Statistical significance in psychological research. *Psychological Bulletin* 70, 3p1 (1968), 151.

[20] Michael C. Frank and Rebecca Saxe. 2012. Teaching replication. *Perspectives on Psychological Science* 7, 6 (2012), 600–604.

[21] Matthew C. Makel, Jonathan A. Plucker, and Boyd Hegarty. 2012. Replications in psychology research: How often do they really occur?*Perspectives on Psychological Science* 7, 6 (2012), 537–542.

[22] J. Sheard, S. Simon, M. Hamilton, and J. Lönnberg. 2009. Analysis of research into the teaching and learning of programming. In *Proceedings of the 5th International Workshop on Computing Education Research Workshop.* ACM, New York, NY, 93–104.

[23] David W. Valentine. 2004. CS educational research: A meta-analysis of SIGCSE technical symposium proceedings. *ACM SIGCSE Bulletin* 36, 1 (2004), 255–259.

[24] Arnold Pears, Stephen Seidman, Crystal Eney, Päivi Kinnunen, and Lauri Malmi. 2005. Constructing a core literature for computing education research. *ACM SIGCSE Bulletin* 37, 4 (2005), 152–161.

[25] Laura Toma and Jan Vahrenhold. 2018. Self-efficacy, cognitive load, and emotional reactions in collaborative algorithms labs—A case study. In *Proceedings of the 2018 ACM Conference on International Computing Education Research.* ACM, New York, NY, 1–10.

[26] Simon, Angela Carbone, Michael de Raadt, Raymond Lister, Margaret Hamilton, and Judy Sheard. 2008. Classifying computing education papers: Process and results. In *Proceedings of the 4th International Workshop on Computing Education Research.* ACM, New York, NY, 161–172.

[27] Lauri Malmi, Judy Sheard, Simon, Roman Bednarik, Juha Helminen, Ari Korhonen, Niko Myller, Juha Sorva, and Ahmad Taherkhani. 2010. Characterizing research in computing education: A preliminary analysis of the literature. In *Proceedings of the 2018 ACM Conference on International Computing Education Research.* ACM, New York, NY, 3–12.

[28] Heiner Evanschitzky, Carsten Baumgarth, Raymond Hubbard, and J. Scott Armstrong. 2007. Replication research's disturbing trend. *Journal of Business Research* 60, 4 (2007), 411–415.

[29] John P. A. Ioannidis. 2005. Why most published research findings are false. *PLoS Medicine* 2, 8 (2005), e124.

[30] Andreas Stefik and Susanna Siebert. 2013. An empirical investigation into programming language syntax. *ACM Transactions on Computing Education* 13, 4 (2013), 19.

[31] Anders Berglund, Mats Daniels, and Arnold Pears. 2006. Qualitative research projects in computing education research: An overview. In *Proceedings of the 8th Australasian Conference on Computing Education—Volume 52*. 25–33.

[32] Laura Krefting. 1991. Rigor in qualitative research: The assessment of trustworthiness. *American Journal of Occupational Therapy* 45, 3 (1991), 214–222.

[33] Denise F. Polit and Cheryl Tatano Beck. 2010. Generalization in quantitative and qualitative research: Myths and strategies. *International Journal of Nursing Studies* 47, 11 (2010), 1451–1458.

[34] SIGCSE Technical Symposium. 2019. SIGCSE 2019—Call For Participation. Retrieved July 24, 2019 from https://sigcse2019.sigcse.org/info/cfp.html#papers.

[35] Brian A. Nosek, Jeffrey R. Spies, and Matt Motyl. 2012. Scientific utopia: II. Restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science* 7, 6 (2012), 615–631.

[36] Matthew T. McBee and Michael S. Matthews. 2014. Welcoming quality in non-significance and replication work, but moving beyond the p-value: Announcing new editorial policies for quantitative research in JOAA. *Journal of Advanced Academics* 25, 2 (2014), 73–87.

[37] Association for Computing Machinery. 2018. Artifact Review and Badging. Retrieved July 24, 2019 from https://www.acm.org/publications/policies/artifact-review-badging.

[38] International Computing Education Research Conference. 2018. ICER'18—Call for Participation. Retrieved July 24, 2019 from https://icer.acm.org/icer-2018/icer18-call-for-participation/.

[39] Alan H. Schoenfeld. 2006. What doesn't work: The challenge and failure of the what works clearinghouse to conduct meaningful reviews of studies of mathematics curricula. *Educational Researcher* 35, 2 (2006), 13–21.

[40] Karl Hunt. 1975. Do we really need more replications? *Psychological Reports* 36, 2 (1975), 587–593.

Received November 2018; revised June 2019; accepted July 2019