Computer Science Education: Guest Editorial

Neil C. C. Brown (King's College London)
Aleata Hubbard Cheuoua (WestEd)
Eva Marinus (Pädagogische Hochschule Schwyz)

ARTICLE HISTORY
Compiled August 3, 2022

DOI

`https://doi.org/10.1080/08993408.2022.2090114`

1. Introduction

Special issues of journals usually focus on a particular research *topic* such as computational thinking or inclusion. This special issue of Computer Science Education is a little different, as the papers cover very different topics, but have in common that they underwent a review process that is new in computer science education: registered reports.

Moreover, all papers in this special issue are (partial) replications of prior work in the field of computer science education. In this editorial we will explain and motivate the importance of these two decisions – focusing on replications and using registered reports – and describe how the review procedure differed from the usual publication process, before giving an overview of the individual papers in this issue.

2. Replications

Replications are an important part of science. Repeating a study can give us increased confidence in the shared findings if they are similar – or reveal mistakes or important contextual nuances in the study if the results differ (Earp and Trafimow, 2015). While replications can increase the reliability of research findings, it is a known problem that replications are not considered prestigious because they are not seen as novel Romero (2019). Moreover, if studies are replicated, the credits seem to go to the authors of the original study and not to the research team that put in the efforts to replicate the study (see, for example, ACM's Artifact Review and Badging Policy ACM (2020)). A survey of computer science education researchers by Ahadi et al. (2016) confirmed that this perception is also present in our field, and that researchers believed replications to be harder to publish than original studies.

Like many other disciplines, only 2% of computer science education publications are replications of previous studies (Hao et al., 2019). If part of the reason for this is systemic – a publication system that makes or conveys that replications are harder to publish – then it seems logical that systemic changes and rewards are needed to increase the amount of replication studies. It is for this reason that the special issue only accepted replications of previous studies, regardless of topic within computer science education. We are pleased that this will lead to a slight increase of the percentage of replications, and we would like to encourage journals and conferences to consider dedicated special issues or tracks specifically for replications of important findings in our field.

3. Registered reports

Peer review of publications is traditionally performed after the study has been conducted, by reviewing a draft of the final manuscript. At this stage, many problems can no longer be addressed – or, if they can be addressed, it is typically a time-consuming endeavour for the researchers. For example, if reviewers want authors to include additional statistical tests or another set of coding categories, then the analysis will need to be re-run. Sometimes reviewers request changes in the data collected during the study that are impossible to fulfill at the final manuscript stage (e.g., asking participants to complete an assessment), leading the study to be rejected. In all, the process seems to be inefficient and ineffective for both authors and reviewers. The logical solution is to plan the reviewing at a time point where much of the study's substance can still be feasibly changed.

Furthermore, restricting reviewing to completed manuscripts is known to lead to publication bias of significant, positive results. In an ideal world, scientists are neutral, unbiased investigators who work towards finding the answer to an important question; the worth of the work should lie in the question being asked, not in the answer that is found. However, reviewers may be tempted to reject studies when there is a negative result: for example in an intervention study where the intervention shows no effect. As a result, it is harder to publish studies with null findings and it can even entice authors to manipulate their study or analysis to show a positive result (Gopalakrishna et al., 2022; Bruton et al., 2020; Grant et al., 2018).

Registered reports aim to solve these problems by changing when a study is reviewed, as shown in figure 1. Authors first write a *stage 1 manuscript* detailing their research plan. This typically looks like the first half of a standard paper: an introduction, a background section or literature review, the research questions and hypotheses and a method section with details of planned data collection and analysis. Crucially, this stage 1 manuscript is reviewed before any data collection is carried out or, in the case of a study with existing data, before any analyses have been conducted.

The stage 1 review carries many benefits. If changes are required to data collection, they can actually be made. If the paper is rejected, much less time is wasted on the part of the authors (and no participant time is wasted). It prevents against publication bias because the accept/reject judgement is made without knowledge of the outcome, thus the decision is only based on the research question being asked and the methods used to investigate it. It also provides transparency: authors cannot later covertly modify their analysis (e.g. to produce a positive result). Such a change would be apparent and would require justification.

If the stage 1 manuscript is approved, authors move forwards to data collection and analysis. Their subsequent complete write-up, now referred to as a *stage 2 manuscript*, is then submitted for approval. In content, a stage 2 manuscript looks much like a classic paper submission. There are two main differences. One difference is that there

Figure 1. The review process for registered reports, where a paper is reviewed after design but before data collection, and then again after the full paper is written. (In classic peer review, only the review labelled "stage 2" is performed.) Image taken from the Center for Open Science (2020), under the Creative Commons Attribution-NoDerivatives 4.0 International License.

will be a description of what changed since stage 1. Modifications are almost inevitable; registered reports do not prevent changes during the research process, they instead make them transparent and justified. For example, some of the papers in this special issue had to change their data collection procedure due to the COVID-19 pandemic. The second difference is that the review is much lighter at this stage, focusing mainly on if the study was conducted as planned and on the presentation of the results, discussion and conclusion.

Registered reports are a relatively new invention. They have been introduced in journals in psychology and education. To our knowledge, this special issue is the first time they have been used in the field of computer science education. They are intended to be an alternative, rather than a complete, replacement for the traditional process: some types of work are not well-suited to registered reports. For example, there remains debate around the compatibility of registered reports, with their review of a prospective data-collection and analysis plan, and qualitative work, which tends to be more iterative and responsive in its approach to collecting and analysing data. However, we believe registered reports can be a valuable contribution to the scientific publication process, as it improves transparency. We are pleased to present this special issue as a proof of concept.

4. The special issue papers

There are four papers in this special issue. All of them replicate at least a part of a previous study, and all were reviewed as stage 1 and as stage 2 manuscript as displayed in Figure 1. We will summarise each of them in turn.

4.1. Replication of "Student Misconceptions of Dynamic Programming", by Shindler et al.

Shindler et al. set out to replicate a 2018 study by Zehra et al. (2018) on how students learn dynamic programming – an algorithms topic often considered difficult. Both studies looked at misconceptions that students have when learning dynamic programming. Shindler et al. increased the sample size compared to the original study and involved multiple institutions, but otherwise broadly followed the same methodology as the original study. Due to the continuing disruption caused by the COVID-19 pandemic, they moved data collection online. The misconceptions found in their study matched those from the original study (such as not recognising the correct recurrence relation), and found more areas of difficulty, such as students inappropriately using a brute force solution, and failing to properly define a base case.

4.2. Unravelling the Numerical and Spatial Underpinnings of Computational Thinking: a Pre-Registered Replication Study, by Finke et al.

Finke et al. aimed to replicate and extend the findings of the validation part of the Computational Thinking test study by Román-González et al. (2017). In this study with Spanish children, Román-González et al. collected normative data for the Computational Thinking test (CTt) and conducted a validation study with a subsample (n=135). For the validation part of the study, they investigated which cognitive skills underpinned performance on the CTt. Finke et al. collected data from 132 Austrian, German-speaking Grade 7 and 8 children (age range 12-15 years). They used a German translation of the CTt and collected data online due to the COVID-19 pandemic. Just like Román-González et al., Finke et al. found that reasoning (assessed with a figural reasoning task) and spatial abilities (assessed with a visualization task) contributed uniquely and significantly to performance on the CTt. However, they also found that additional variance was explained by complex numerical abilities (i.e., algebraic skills). Finally, Finke et al. replicated the finding that boys outperform girls on the CTt. However, in contrast to Román-González et al., they found that these differences were largely independent of the gender differences in cognitive skills.

4.3. Assessing Individual Contributions to Software Engineering Projects: A Replication Study, by Hundhausen et al.

Motivated by the need to examine objective measures of individual contributions to team projects, Hundhausen et al. performed a replication of *Assessing Individual Contributions to Software Engineering Projects with Git Logs and User Stories* (Buffardi, 2020). They expand upon Buffardi's work by including multiple institutions which varied in the software engineering courses offered. Hundhausen et al. largely replicated the data sources and measures used by Buffardi, but with some variations (e.g., to account for how each participating course assigned grades, collecting demographic data on student participants). They were able to replicate four of five significant findings from Buffardi's study; *relative commit shares* (i.e., number of GitHub commits per member team relative to the expected number of commits per team member) was found not to be a predictor of peer contribution ratings. Furthermore, they found more significant associations between subjective and objective metrics of individual contribution to team projects than Buffardi did in the original study. This replication, in addition to providing support and new insight into Buffardi's results, highlights the importance of including a sufficient number of participants to detect effects and attending to variation across study sites that can influence data measures.

4.4. Reevaluating the Relationship between Explaining, Tracing, and Writing Skills in CS1 in a Replication Study, by Fowler et al.

Fowler et al. performed a study to replicate a slightly simplified hierarchy of skills from a paper by Lopez et al. (2008). The hierarchy concerns the dependencies between the skills of reading, tracing and writing program code during introductory programming lessons. Fowler et al. expanded on the original study by using a larger sample size and investigating multiple possible structural equation models to see which could best explain the data. They found that although the original hierarchy from Lopez et al. did not appear among the best models, similar models did appear. The authors discuss the limitations of this approach, noting that it can only reveal correlational relationships between the skills and cannot directly inform the fundamental question: in what order should these skills be taught. They propose alternative future study designs that could address this question.

References

Artifact Review and Badging - Current, Aug. 2020. URL https://www.acm.org/publications/policies/artifact-review-and-badging-current.

A. Ahadi, A. Hellas, P. Ihantola, A. Korhonen, and A. Petersen. Replication in computing education research: Researcher attitudes and experiences. In *Proceedings of the 16th Koli Calling International Conference on Computing Education Research*, Koli Calling '16, page 2–11, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450347709. . URL https://doi.org/10.1145/2999541.2999554.

S. V. Bruton, M. Medlin, M. Brown, and D. F. Sacco. Personal motivations and systemic incentives: Scientists on questionable research practices. *Science and Engineering Ethics*, 26(3):1531–1547, Jun 2020. ISSN 1471-5546. . URL https://doi.org/10.1007/s11948-020-00182-9.

K. Buffardi. *Assessing Individual Contributions to Software Engineering Projects with Git Logs and User Stories*, page 650–656. Association for Computing Machinery, New York, NY, USA, 2020. ISBN 9781450367936. URL https://doi.org/10.1145/3328778.3366948.

Center for Open Science. Registered reports, 2020. URL https://www.cos.io/initiatives/registered-reports.

B. D. Earp and D. Trafimow. Replication, falsification, and the crisis of confidence in social psychology. *Frontiers in Psychology*, 6, 2015. ISSN 1664-1078. . URL https://www.frontiersin.org/article/10.3389/fpsyg.2015.00621.

G. Gopalakrishna, G. ter Riet, G. Vink, I. Stoop, J. M. Wicherts, and L. M. Bouter. Prevalence of questionable research practices, research misconduct and their potential explanatory factors: A survey among academic researchers in the netherlands. *PLOS ONE*, 17(2):1–16, 02 2022. . URL https://doi.org/10.1371/journal.pone.0263023.

D. B. Grant, G. Kovács, and K. Spens. Questionable research practices in academia: antecedents and consequences. *European Business Review*, 30(2):101–127, Jan 2018. ISSN 0955-534X. . URL https://doi.org/10.1108/EBR-12-2016-0155.

Q. Hao, D. H. Smith IV, N. Iriumi, M. Tsikerdekis, and A. J. Ko. A systematic investigation of replications in computing education research. *ACM Trans. Comput. Educ.*, 19(4), aug 2019. . URL https://doi.org/10.1145/3345328.

M. Lopez, J. Whalley, P. Robbins, and R. Lister. Relationships between reading, tracing and writing skills in introductory programming. In *Proceedings of the Fourth International Workshop on Computing Education Research*, ICER '08, page 101–112, New York, NY, USA, 2008. Association for Computing Machinery. ISBN 9781605582160. . URL https://doi.org/10.1145/1404520.1404531.

F. Romero. Philosophy of science and the replicability crisis. *Philosophy Compass*, 14(11):e12633, 2019. . URL `https://compass.onlinelibrary.wiley.com/doi/abs/10.1111/phc3.12633`. e12633 PHCO-1228.R1.

M. Román-González, J.-C. Pérez-González, and C. Jiménez-Fernández. Which cognitive abilities underlie computational thinking? criterion validity of the computational thinking test. *Computers in Human Behavior*, 72:678–691, 2017. ISSN 0747-5632. . URL `https://www.sciencedirect.com/science/article/pii/S0747563216306185`.

S. Zehra, A. Ramanathan, L. Y. Zhang, and D. Zingaro. Student misconceptions of dynamic programming. In *Proceedings of the 49th ACM Technical Symposium on Computer Science Education*, SIGCSE '18, page 556–561, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450351034. . URL `https://doi.org/10.1145/3159450.3159528`.