Launching Registered Report Replications in Computer Science Education Research

Neil C. C. Brown
King's College London
London, UK
neil.c.c.brown@kcl.ac.uk

Eva Marinus
Pädagogische Hochschule Schwyz
Goldau, Switzerland
eva.marinus@phsz.ch

Aleata Hubbard Cheuoua
WestEd
San Francisco, CA, USA
ahubbar@wested.org

ABSTRACT

Background and Context. The quality of a research field is greatly determined by the quality of its publications. However, in many research fields this quality is endangered because of biases towards publishing positive results, which have led to questionable research practices that negatively impact the reliability of the reported findings.

Objectives. We explored two approaches that could guard against these questionable research practices for computer science education research: 1. The replication of prior studies to confirm (or refute) previous findings and 2. requiring authors to submit a registered report, containing hypotheses, methods, and analytic procedures, before conducting the studies.

Method. Over the span of 18 months, we organized a special issue of the Computer Science Education journal that only accepted registered reports that replicated a previous computer science education research study. Registered reports involve peer review and approval at the research design stage, before data collection begins. The editorial process was thus modified to accommodate multiple rounds of review and a longer time period between original and final submissions. We believe this is the first use of registered reports in computer science education research. A questionnaire gathering feedback on the new process was also administered to the authors of the accepted reports.

Findings. We found seven author teams willing to submit a manuscript for the special issue. Out of this pool, preregistered reports from five teams were accepted to be taken forwards. One team then withdrew because the ethics procedure at their institution exceeded the special issue timeline. The remaining four author teams conducted their studies and resubmitted full papers that were accepted, pending some final corrections. Authors' feedback about the registered reports process was positive.

Implications. We demonstrated that it is possible to attract interest for registered report replications in a computer science education journal issue, and to successfully conduct the necessary editorial steps. Future efforts should attend to challenges related to modified research and journal submission timelines and consider adding a second review cycle for the first stage of registered reports. While the procedures used in this special issue may be suitable for many research approaches, further discussion is warranted on how they can be combined with exploratory research (as opposed to hypothesis testing research) and how they can be adapted to non-positivist research.

ICER 2022, August 7–11, 2022, Lugano and Virtual Event, Switzerland

ACM ISBN 978-1-4503-9194-8/22/08…$15.00
https://doi.org/10.1145/3501385.3543971

CCS CONCEPTS

• **Social and professional topics** → **Computer science education**; • **General and reference** → *Evaluation.*

KEYWORDS

Replication, Preregistration, Registered Reports, Metascience

ACM Reference Format:
Neil C. C. Brown, Eva Marinus, and Aleata Hubbard Cheuoua. 2022. Launching Registered Report Replications in Computer Science Education Research. In *Proceedings of the 2022 ACM Conference on International Computing Education Research V.1 (ICER 2022), August 7–11, 2022, Lugano and Virtual Event, Switzerland.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3501385.3543971

1 INTRODUCTION

Much of the value of computer science education research (CSER) lies in increasing our understanding of computer science education in order to make real impact in areas of concern, for example improving student learning or increasing diversity. For this to be possible, it is important that CSER is high quality and that its findings are reliable, so that stakeholders – such as teachers, students, funders, national agencies – can confidently act on its findings. There are many cautionary tales of poor quality research from other scientific disciplines that it would be wise to understand and to avoid as CSER grows.

One popular example is the replication crisis, which was first reported in psychology. Since 2010, many studies in this research field have come under strong scrutiny because their key research findings, which have been passed on and taught to later generations of scholars for decades, were found not to replicate [17]. Since then, fears of such a replication crisis have spread to a wide variety of other fields such as economics [49], health informatics [16] – as well as education [73] and computer science [15], the two neighbouring fields of CSER.

The serious consequences of having a large amount of publications with unreliable results in a field of research are shown in the work of Lortie-Forgues and Inglis [46]. These researchers reviewed preregistered, large-scale randomised controlled trials – often thought to be a research gold standard – in education and found that most still reported uninformative findings. They posit that one explanation is that although these studies themselves are

high quality, "many of the interventions studied are ineffective because the literature on which they are based is unreliable". It seems that in research, a lowering tide sinks all boats: shaky research foundations cannot be overcome by simply running a new high quality trial if it relies on those foundations.

Unreliable research can be hard to retract and can lower stakeholders' confidence in other findings. For example, the idea of learning styles has become entrenched in education folklore [53], even though the evidence is unconvincing. As Cuevas [20] describes: "a substantial divide continues to exist, with learning styles instruction enjoying broad acceptance in practice, but the majority of research evidence suggesting that it has no benefit to student learning, deepening questions about its validity." Even if educators were to be convinced that learning styles are unsupported by evidence, how could they be confident that the next suggestion from education researchers would not turn out the same way? Hovey et al. [39] suggest that CS educators already dismiss research: "Evidence from empirical [education] studies [for the purpose selecting a new teaching practice] was not typically a consideration, and in at least one case, a faculty member dismissed evaluative studies as being inherently biased."

One explanation for the unreliability of some research is that undesirable incentives lead to what have been termed Questionable Research Practices (QRPs) [42]. With prestige and rewards (e.g., career promotion) attached to widely-cited publications, authors are under pressure to get their papers published. The publication process itself is biased towards novel results [61], and statistically significant results are more likely to be cited [7, 54]. Indeed, Serra-Garcia and Gneezy [64] found that papers which later fail to replicate are more likely to be cited than those which do replicate, even after the replication failure has been published – suggesting that novelty is more important for citations than reliability.

With novel and positive results thus encouraged, researchers may turn to QRPs to produce them. The well-known QRP of p-hacking (i.e., continuing data collection or selecting a set of data to produce a statistically significant result) has been found to be widespread in many scientific fields [36, 42, 56]. Peterson [58] conducted an eye-opening study on one research lab in infant cognition, in which many QRPs were observed, including HARKing (Hypothesising After Results are Known) – which was found in the study to be common practice:

> "The structure of these meetings was similar across labs. A professor or graduate student would e-mail a short document to the lab a few days before and then hand out those same pages at the beginning of the meeting. Usually, they would contain a couple of box plots or bar charts. The experimenter would then point out where statistical significance was reached and then ask the lab for help figuring out what could be argued from the results. The lab would attempt to collectively craft a story out of the significant findings. When a clear and interesting story could be told about significant findings, the original motivation was often abandoned." [58, pg. 6]

It would be naïve to assume that CSER is somehow immune to these pressures and to the QRPs that tend to follow. Ahadi et al. [1] surveyed computing education researchers and found that "researchers in our field hold many of the same biases as those in other fields experiencing a crisis in replication. Furthermore, while the respondents agree that published work should be verifiable, they doubt this standard is widely met in the computing education field." We believe it is unproductive to phrase this as a moral issue – viewing those committing QRPs as bad people or sinful, or thinking that the problem is limited to a few malicious actors. We should recognise that QRPs are incentivised by the system [9, 33], and that we are all liable to temptation – thus it is important to have checks in place for all of us (including the authors of this article!). Researchers will engage less in QRPs if they think it will be detected by reviewers [32], or if they are incentivised to avoid QRPs [9].

Although other disciplines offer cautionary tales for CSER, they also offer potential solutions to the problems. This includes both *detection* of the problems described above – poor replicability, QRPs – and their *prevention*. There are two main aspects of guarding against a replication crisis: first, there must be sufficient replications to detect such problems, and second, there must be good measures in place to make sure that the published research findings are likely to replicate. One effective way to enhance replicability of research findings is to request authors to submit a registered report [12] before conducting the study. With a registered report, peer review is not only conducted after the study is completed, but also after the research planning and design stage, crucially *before* data collection and analysis. As a result design issues can be correctly addressed, and the planned analysis is determined and externally recorded before the data is collected. The preliminary accept/reject decision is made at this stage without knowing the study's results, and thus cannot be dependent on the outcome of the analysis.

This article reports on a special issue of the Computer Science Education journal where the submissions had to be replications and were reviewed in a registered reports style: an attempt to try both interventions at once. The contributions of this paper are:

- An explanation of replications and registered reports, and the issues with the scientific process that they can solve, with particular reference to computer science education research (section 2).
- A case study report on how these two aspects were implemented for computer science education research (section 3), with reflections on what worked well and what did not (section 4).
- Recommendations for computer science education journals and conferences on how to implement changes to better support replications and registered reports, as well as discussion of the limits of these techniques (section 5), including with respect to non-positivist work (subsection 5.3).

There have been opportunities in CSER for informal early feedback like ICER's work-in-progress workshops, and occasional voluntarily pre-registered studies in the field of computer science education (e.g., Marinus et al. [50] and Brown et al. [8]) – that is, a study that has registered their research questions, design and analyses in advance without peer review on a platform like the Open Science Framework. However, we do not know of any computer science education studies where the research plan was formally peer reviewed by the publication venue as is the case with registered reports, and we believe this article reports on the first use of registered reports in computer science education.

2 BACKGROUND

The special issue reported on in this paper focused on two interventions to prevent and/or detect questionable research practices: requiring submissions to be replications of prior work, and reviewing the submissions in a registered report style. We will discuss the motivation for replications and registered reports in more detail, in turn, after first introducing and explaining the questionable research practices that they help to prevent.

2.1 Questionable research practices

Research is a field that is built on trust. If a researcher is not honest and transparent, it is difficult to detect this during peer review. Outright data fabrication appears to be rare [31], although cases usually take a long time to discover [19], because peer reviewers do not have sufficient information available to detect fabrication – thus cases usually result from concerns by collaborators instead [19]. Much more prevalent are questionable research practices (QRPs): minor alterations to the data collection or analysis procedure to produce desired results [30, 42]. Peer reviewers in the current system also lack sufficient information to reliably detect these cases, and thus are forced to trust (or at least cannot falsify) the claimed procedure in the submitted manuscript. Some examples of questionable research practices are as follows (see John et al. [42] for an extensive list):

- **Hypothesising After Results are Known (HARKing)** is the practice of changing the claimed investigation to match what the data showed. As one researcher summed up the practice in their own laboratory (as reported by Peterson [58]): "You want to know how it works? We have a bunch of half-baked ideas. We run a bunch of experiments. Whatever data we get, we pretend that's what we were looking for." The reason why this is a problem is not necessarily obvious: the data did show the effect, after all. But the data will always show evidence for some *post hoc* imagined hypothesis; by pretending that was the aim all along, other researchers will trust or investigate a pattern that might well have been mere random noise in the original data.
- **Publication bias** refers to the preference among reviewers and/or authors for positive results (e.g. finding a result which confirms the experimental hypothesis). Publication venues may reject work that does not have a positive result, or results that are not considered novel. Authors may have the same biases when deciding whether to write-up work for publication (partly as a result of the expected bias of the venues). Such a bias means that only positive results are shared, and thus everything investigated seems to have an effect, distorting our knowledge of the world because we do not see investigations (original or replications) that do not show an effect. With publication bias, science becomes a search for effects, rather than a neutral investigation of the world.
- **p-hacking** is the practice of deliberately manipulating statistical tests in order to produce a significant result. Quantitative work commonly uses frequentist statistical tests, where usually a test is considered statistically significant if the resulting p-value of a test is less than or equal to 0.05. The practice of p-hacking interacts with publication bias [28] – it is only necessary to p-hack to get published if publication is biased towards positive results. Kühberger et al. [44] performed an analysis which suggested that in psychology there are an unexpectedly large number of p-values just under the 0.05 threshold – a sign of p-hacking. There are several sub-types of p-hacking, including the following examples which are all hard to detect during peer review:
  - **Selective stopping** is the practice of continuing data collection while results are not significant, and stopping at the exact point when they happen to be significant. It is difficult for a reviewer to know that this was the real reason the data collection ended.
  - **Data fishing** is the practice of performing multiple tests on the data to find one that is significant, and then either not reporting the non-significant tests or not correcting for the increased chance of false positives.
  - **Analysis modification** is the practice of changing the tests at analysis time to produce a significant result. The results of these tests depend on decisions taken by the researcher, such as which outliers to remove or retain, or exactly which factors to include in a statistical model. These analysis choices are subtle and often potentially valid [29]. It can be difficult for reviewers to distinguish between reasonable decisions that coincidentally produce a significant result, from decisions made to deliberately produce a significant result.

To give some concrete CSER examples of these QRPs, imagine a study that investigated whether using Python rather than C++ improves retention rates, but the statistical test showed that there was no difference in retention. Publication bias would mean the study would be rejected (or not even submitted) for publication if reported as-is. p-hacking could mean excluding students with prior experience of C++ programming, if that resulted in a remaining subset with a significant difference. The researcher may notice that although choice of programming language made no difference, students who reported using an IDE had higher retention rate regardless of language, and thus HARK to write-up a study about how IDE use led to higher retention (across two languages, no less), without mention of the original aim to separate by language. The difficulty of detecting these as a peer reviewer is that avoiding HARKing requires disputing (without evidence) the authors' claimed study aim, the p-hacking may be the result of a reasonable *a priori* decision rather than a *post hoc* search for significance, and the publication bias may cause the paper to never be written, or cause it to be rejected by other reviewers.

Some estimates of the prevalence of questionable research practices are as high as one-third of scientists [23] to over one half [32, 42] who engage in such practices (although see the Fiedler and Schwarz [24] rebuttal of John et al. [42] which suggests this is a flawed way to measure). A recent meta-analysis by Xie et al. [74] suggests a prevalence of 12.5%. George [30] summarises older studies (1987–2008) which suggest a much lower prevalence rate of 2% although the author notes that these self-admission rates may be quite inaccurate (and understanding of QRPs may have been lower in the past). Nevertheless, George goes on to state:

> "It is arguable that in aggregate more damage is caused by the less serious forms of questionable research practices and from sloppiness or incompetence than

> from data fraud – largely because these other sources of data errors are more common." [30, pg. 17]

In a survey of research participants, Bottesini et al. [6] found that "between 68% and 81% of participants reported that p-hacking, [not publishing negative results], HARKing, and fraud are not acceptable, ... [and] that replication is preferable to moving on without replicating", and suggest that may be unethical towards participants to commit such practices with the data they contribute.

The cause of these questionable research practices is thought to be largely structural incentives [9, 33]: novel positive results are more likely to be published, and more likely to attract funding. Researchers feel pressure to gain publications [32] and are thus more likely to engage in QRPs. We posit that such structural incentives can only be tackled via structural changes, rather than relying solely on individual morality or ethics – a view that Bruton et al. [9] found is shared by other researchers.

In this paper we will focus on two ways to potentially resolve the issues with questionable research practices:

- **replicating** studies to gain confidence in the findings, and
- changing the publication process to prevent HARKing and publication bias via the use of **registered reports**.

We will explain each approach in turn.

2.2 Replications

Replication has long been proclaimed as an important part of science. In his oft-cited book, "The logic of scientific discovery" [60], Popper wrote:

> "We do not take even our own observations quite seriously, or accept them as scientific observations, until we have repeated and tested them. Only by such repetitions can we convince ourselves that we are not dealing with a mere isolated 'coincidence'…" [60, pg. 23]

The underlying principle is that if we can replicate a study, we become more confident in the shared findings. The power of a replication lies not just in the increase in data that it brings, but also the subtle variations in contexts and researcher decisions across the multiple independent attempts to conduct the same experiment. Perhaps the first study contained a mistake in the data or analysis, perhaps it was very tightly bound to its context, or maybe it was sheer coincidence; we cannot be sure if any of these apply if the study is never repeated.

Earp and Trafimow [22] stress that if a replication produces a different result than the original study, it is incorrect to think that the first study is definitively falsified: there are instead a wide range of possibilities. The first *or* second study could be incorrect, or a difference in the study contexts may ameliorate the effect. The outcome of a replication attempt should thus be considered informative, rather than a definitive yes/no determination. However, if two studies produce different results (often referred to as a failure to replicate), it is a sign that the original result *may* not be reliable or generalisable.

A well-known example in CSER of a failure to replicate is the preprint paper "the camel has two humps" [21] by Dehnadi and Bornat which claimed to have found a test that could determine who would succeed at programming and who would not. Replications by Caspersen et al. [11] and Lung et al. [47] as well as a team including Dehnadi and Bornat (Bornat et al. [5]) did not find the same effects. Bornat later retracted many of the original paper's claims, in light of this evidence [4].

Similar to most other disciplines, only 2% of computer science education papers are replications [34]. The cause of the lack of replications is well-known: a desire for novel knowledge. As Romero [61] describes:

> "A central component of the reward system of science is the priority rule, that is, the practice of rewarding only the first scientist that makes a discovery. This reward system discourages replication." [61, pg. 5]

In a survey of computing education researchers by Ahadi et al. [1], the respondents agreed that novel works are considered more prestigious than replications in CSER. To correct this problem, we must therefore alter the scientific reward system to encourage and value replications. It is for this reason that the special issue only accepted replication studies.

2.2.1 Types of replication. There is a general tension in the concept of replications surrounding possible context differences. A perfect replication is uninformative. Computer science is perhaps one of the few areas where this is possible: if we run the same deterministic program twice and get the same result, we have learned nothing. In reality, perfect replications are rarely possible, especially with human participants. If you run an experiment to see if people prefer red or blue, you may not get the same result even if you run the experiment twice with the same participants, let alone a different set of participants, perhaps in a different country.

A replication by new researchers with a very similar context is generally very informative, for example repeating the same procedure but in a different laboratory with a different (but demographically similar) set of participants. Whereas a study investigating the same aim with totally different methods and context would not typically be deemed a replication. There is clearly a sliding scale in-between as to what is considered a replication and what is not. As Chhin et al. [13] noted, there are "fifty shades of replication" with scholars using differing terms and definitions to describe replication approaches (not to mention whether the replication is carried out by the same researchers or new researchers). Researchers often distinguish between exact replications, empirical replications and conceptual replications as gradations on this scale. *Exact replications* examine the same (or highly equivalent) participants/artifacts using the same procedures as the original study. *Empirical replications* examine different populations using the same materials and procedures as in the original study. *Conceptual replications* examine the same phenomenon/hypothesis using different procedures than the original study. To further clarify these definitions, we provide examples of prototypical CSER studies and how each might be replicated with these three approaches in Table 1.

These types of replications are primarily centred around the [post-]positivist philosophy of science on which most quantitative research builds. Qualitative research typically adopts a different philosophy of science (e.g., constructivist rather than [post-]positivist) and as Talkad Sukumar and Metoyer [69] explain, its findings are usually not intended to generalise directly:

Prototypical Study	Replication Studies
Secondary students participating in an informal learning program are interviewed about their interest in computing before and after the 6-week program	• *Exact:* Researchers set up the same program with an equivalent population of students to see if the effect replicates. • *Empirical:* Researchers use the same interview protocol and analytic approach but with a program in a different country. • *Conceptual:* Researchers run the same learning program but use different measures to investigate interest in computing.
Several studies have proposed ways to improve student performance on the Rainfall Problem (a programming task originated by Soloway [68]).	• *Exact:* Researchers use one of the suggested approaches in a classroom with a similar population of students. • *Empirical:* Researchers use several of the techniques in a randomised control trial to see which offers the most improvement. • *Conceptual:* Researchers use one of the suggested approaches in a lesson on another topic many students find difficult.
The most common textbooks used in introductory college computing courses were examined using a content analysis approach to categorize the nature of their programming examples	• *Exact:* Researchers re-examine the same textbooks using the same content analysis approach. • *Empirical:* Researchers apply the same approach to textbooks aimed at introductory high school computing. • *Conceptual:* Researchers apply machine learning to categorize the textbooks.

Table 1: Examples of Replication Studies

> "The positivist perspective assumes that reality or knowledge exists as the objective truth independent of those who study it while the constructivist perspective views knowledge as subjective, constructed through interaction and inseparable from those who study it. Hence findings of quantitative studies can be replicated and generalized to larger populations whereas qualitative findings only hold for and describe the individual cases." [69, pg. 1]

They go on to propose that qualitative replications "encompass repeating some aspect of an earlier study's design and a subsequent interpretive comparison." Tuval-Mashiach [72] concurs and suggests that therefore *conceptual* replications are applicable and valuable in qualitative research, rather than exact or empirical replications which may not be applicable.

2.2.2 Importance of replications. A practical example of why replications are important – and how a lack of them results in fragile knowledge – is that of Daniel Kahneman's book "Thinking, fast and slow" [43]. It was a book written by Nobel prize winner Kahneman to summarise other researchers' results, primarily from social psychology about human behaviour. It later transpired that many of these effects did not replicate and appeared to be spurious. In online discussion with other researchers, Kahneman (admirably) accepted that many of the cited studies did not replicate, thus undermining some of the statements made in his book [51]. Another example arises from gene research, where many small-scale studies found significant effects of a gene variant 5HTTLPR, but a later study with a much larger sample size (hundreds of thousands of people) found that no such effects replicated in the larger sample [3].

Much of the recent discussion around replications was triggered by the large-scale reproducibility project, which set out to replicate a set of 100 findings from psychology [17] and found that one-third to one-half of them replicated. The reader may form their own speculation what the proportion may be for computer science education research; in the Ahadi et al. [1] survey, computer science education researchers estimated the proportion to be 30%. It is clear that in order to *determine* this value, we must conduct more replications, and that in order to *increase* this value, we must look for techniques that can increase the reliability of published research – such as registered reports.

2.3 Registered reports

In the fields of psychology and education many journals now offer the possibility to submit registered reports. In a registered report, authors must first submit a "stage 1" research plan describing their research topic, research questions, methods and analysis plan for review *before* the data collection and analyses take place. This plan is then peer reviewed and a provisional decision is made to accept or reject the paper. If the registered report is approved, the paper will be accepted for publication, irrespective of the outcomes, as long as the researchers conduct the study as planned. This is checked in a second round of review on the full "stage 2" submission (see Figure 1 for an overview of the modified review process). Registered reports are not a constricting straitjacket. Changes can still be made after the initial review if appropriate (e.g. moving data collection online due to a pandemic, changing recruitment strategy after difficulty recruiting, adjusting analysis due to missing data) but in a registered report they become transparent to the reviewers of the final paper, and must be justified. For example, in the preregistered study by Bottesini et al. [6], they found that in the preregistered analysis, a higher proportion than expected approved of outright fraud, so they conducted an exploratory extra analysis with tighter exclusion criteria, which suggested this may have been caused by some participants not answering seriously. Both the preregistered analysis and exploratory analysis were reported, labelled and justified appropriately.

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 [26], under the Creative Commons Attribution-NoDerivatives 4.0 International License.

This approach takes away several reasons why studies fail to replicate:

- First, the stage 1 review will allow correction of poor quality study design caused by inexperience or a lack of planning in the initial stages of the research. For example, a specific control group could be missing, there might be better measurement instruments than the authors selected or the number of suggested participants might be too low. The final point relates to the issue that studies are often under-powered. This means that their sample sizes are too small to detect the target effects, and, in the worst case, that the outcomes of the statistical tests are almost at random. Reviewers of a registered report can suggest authors address this problem by conducting a power analysis, which will directly inform what sample size is required to conduct the analyses to detect the expected effect size.
- Second, because researchers are assured of a publication when they conduct their studies according to the approved research plan, it takes away potential bias in both authors and reviewers towards wanting to report positive (i.e. statistically significant) results, which can encourage researchers to conduct questionable research practices like HARKing and p-hacking as described in subsection 2.1. The accept/reject decision is made before data collection so it cannot be influenced by the analysis outcome – reviewers can no longer reject the paper because the analysis "did not find anything" (i.e., did not reject the null hypothesis).

In sum, registering research questions, hypotheses and the corresponding analyses in advance is a very powerful measure to prevent the publication of unreliable, hard-to-replicate results. A recent study in psychology found that registered reports had 44% positive results, in contrast to the 96% positive results found in matched standard publications in the same time period [63]. This suggests that registered reports can have a significant impact on the issues of publication bias and Type I errors.

Although registered reports are a good tool to counter replication issues, researchers may question whether these measures are too restricting and if they are suitable for all research designs and approaches. For example, consider exploratory or data-driven research where it may be very hard or even impossible to define research questions, hypotheses and an analysis plan beforehand. In addition, while researchers may have specific hypotheses about some parts of their research plan, they may not be sure about the other aspects. Registered reports cater for this by allowing researchers to make a distinction between these two aspects of their research by explicitly labelling the latter as exploratory. This distinction between the registered and exploratory parts of the study is also carried over to the final manuscript, so that readers are aware which results will be more likely to generalise and hold in the future (those from planned analyses) and which findings may need more data and replication (those from exploratory analyses). How registered reports could be implemented for non-positivist studies is discussed later in the paper, in subsection 5.3.

3 A CASE STUDY OF REPLICATIONS AND REGISTERED REPORTS IN CSER

We collaborated as co-editors on a special issue of the Computer Science Education journal focused on registered report replications in computer science education research. The idea for the issue was born out of conversations between the journal editors and the first author around avoiding the replication crisis in CSER. The first author then assembled an editorial team to represent prior experience in registered reports, interest in replication studies, and a range of expertise in quantitative and qualitative approaches to CSER. In this section, we report on our special issue process and the modifications in both the research and review processes (see the overview in Figure 1) compared to the standard review process.

3.1 Research process

In November 2020, a call was distributed soliciting *Stage 1* papers for the special issue. These initial papers were prepared after authors identified a study to replicate but before beginning any data collection. Stage 1 submissions required:

- an introduction to justify the replication;
- a literature review containing a summary of the relevant literature, especially the literature published since the publication of the original paper;
- a methods section detailing the type of replication to be implemented (see Table 1);
- a plan outlining the analysis approach including any additional analyses that went beyond those in the original study; and
- risks to completing the study within our proposed timeline (e.g., pending funding, status of ethics application).

We formed policies related to several expected questions from prospective authors:

Can proposers go beyond the original study? We expected some authors would propose additions to improve or enhance the original studies they replicated such as including a larger sample size (as justified by a power analysis or theoretically-driven rationale), an additional training condition, an additional participant group, additional scoring categories, or additional tests to assess the skills of the participants. The authors were free to conduct additional analyses as long as they were outlined in the Stage 1 paper and labeled as such. To enable direct comparisons between the original and replication studies, we asked authors to explain in the introduction how their replication would add further information to the claims and outcomes of the original study. For both exact and empirical replications, authors needed to conduct an as-close-as-possible replication of the original study before examining and reporting the additional analyses.

Who should be involved in replications? In planning the special issue, we made two decisions regarding how original authors could be involved in the new studies. First, we did not allow authors of original studies to replicate their own work. This decision was driven by our desire to encourage independent replication of past studies. Second, we did not require or recommend proposers to include the original author on their paper. We felt this could lead to undue pressure to produce results favorable to the original study. However, we encouraged proposers to contact the authors of the original paper to request sharing the materials, coding schemes, analysis scripts and other useful information to support the replication.

What amount of flexibility should be allowed between Stage 1 and Stage 2? Educational research is inherently messy and study plans often change for a variety of factors. As such, we did not want to prevent authors from making reasonable modifications of Stage 1 study designs. For example, we considered revising data collection plans to conduct observations virtually instead of in-person due to pandemic restrictions a reasonable change to make after stage 1 submissions. Following the goals of registered reports, we wanted authors to be transparent about any changes that arose and to justify their decisions in the stage 2 manuscript. So, we decided to accept stage 2 papers that made reasonable adjustments and documented and justified these differences. After all, the aim is to increase transparency, not to add unreasonable restrictions.

3.2 Peer Review Process

Given the rarity of replication studies in CSER and differences in reviewing a registered report as compared to a full manuscript, we wanted to expand the typical journal reviewer pool and provide additional supports for reviewers. A call for reviewers with experience in either replications or registered reports was released simultaneously with the call for papers. We also developed a 30-minute training for reviewers that covered most of the information in section 1 and section 2 of this paper as well as modifications that we made to the journal's standard reviewing form. For example, where the standard review form asked *"Are the methods appropriate to the problem; do they provide sufficient evidence and data to back up their claims?"*, we included the following additional prompts:

- Do the authors define the type of replication that they will do (exact, empirical, conceptual) or is it at least clear which one will be done?
- Do they clearly distinguish between the materials/approaches of the original study and the additions of the replication study?
- Are there any potential confounding factors in the research design (that is: will the planned data collection answer the question, or could there be an additional factor that could explain the outcomes). If so, do the authors address this?
- When recognising design flaws, please add advice on how to change the design when possible.
- Do the authors correctly summarize and interpret the original analyses?
- Do they explain how any newly planned additional analyses will be done?
- When a quantitative replication: have they included a power analysis?
- Are there any potential holes in the analysis plan? For example: what if the raters disagree, what if the randomly allocated groups are mismatched, etc.

There were two major rounds of reviewing: a peer review for the stage 1 manuscript and an editor review for the stage 2 manuscript. The rationale for the latter was that the research design had been reviewed by subject-matter experts in the first round and the second round was a matter of checking that the analysis matched the plan, and that the study had been written clearly. This felt more like an editor's job than a peer reviewer's job, and it would be more expedient to not involve external reviewers again. We discuss the advantages and disadvantages of this decision further in section 4.

3.3 Timeline

The full process took around 18 months from initial submission of stage 1 manuscripts to submission of the final papers (see Table 2). It is important to remember this is not comparable to the time taken for a standard submission, because this 18 month period includes the finalisation of the research design, and the entirety of the data collection and analysis. In contrast, the timeline for a standard submission would only include revision of the final manuscript after all other aspects of the research had been completed. To give examples from CSER journals: ACM's TOCE state they aim to give a recommendation within 90 days of authors submitting the [full] paper, and T&F's CSE currently state a 96 day average "from submission to first post-review decision." This is comparable to the initial stage 1 review duration in our timeline.

We believe that registered reports may actually be faster overall than traditional submissions in many cases. The stage 2 review is unlikely to request significant rewrites or extra data collection, because any substantive issues over the design will have been dealt with at stage 1, before the majority of the paper is written. Additionally, the time to publish standard articles can include a hidden extra delay: that of resubmitting to another journal in cases where the submission is rejected by one journal (potentially after multiple rounds of review). With registered reports, the preliminary decision to accept is made before data collection, so such a rejection wastes less time for authors (and participants, who will not need to donate their time if the study is not run).

Date	Stage
Nov 2020	Initial submission of stage 1 manuscripts
Feb 2021	Peer reviews of stage 1 manuscripts returned to authors
Jan 2022	Submissions of stage 2 manuscripts
Feb 2022	Editor reviews of stage 2 manuscripts returned to authors
Apr 2022	Final submission of stage 2 manuscripts

Table 2: Special Issue Timeline

3.4 Materials

To encourage replication of our own process of running the special issue, we have created an Open Science Foundation repository (https://osf.io/thq3p/) that includes: the call for authors and reviewers for the special issue, the slides of the training presentation given to reviewers, the annotated expanded review form for stage 1 reviewing, and the evaluation questionnaire that was given to authors at the end of the process.

3.5 Outcome

The final submissions for the special issue have now been received, and the process has been largely successful. Authors from three (of the final four) papers filled in an optional evaluation survey (conducted in accordance with the Research Ethics Committee process of King's College London), and their responses inform this section.

3.5.1 Acceptance and timescales. Seven stage 1 submissions were accepted, six were sent out for review, and five were accepted to proceed. One of the five papers did not receive institutional ethical approval within the timescale and thus only four papers returned their stage 2 manuscripts. One author who did return a stage 2 manuscript noted that the timescales were tight for the data collection and analysis. Taken together, this suggests that the timescale for the special issue of 11 months for data collection, analysis and write-up was too short for several authors. Some authors were constrained for time by external constraints on when data collection could occur, for example only during teaching terms or only at exam time.

3.5.2 Replications. One author commented that it was useful to permit replications that were not only exact replications. Another author pointed out that with computing being a changing field, close replication may not be appropriate for technical reasons: some old technologies (e.g. old programming languages) may not be suitable or even available for replication. Additionally, some authors had to adjust for the effects of the COVID-19 pandemic, which meant in-person data collection was not feasible, which made a difference to the research design, given that the original studies being replicated may have been conducted in-person.

3.5.3 Allocating the reviewers to the papers. Our call for reviewers willing to be trained in registered reports resulted in an enthusiastic and knowledgeable set of reviewers. However, we faced some challenges in matching the expertise of the reviewers to the topics and analysis methods of the papers, and we separately recruited some extra reviewers as subject experts, who had not attended the registered report training. We subsequently shared all our training materials with these new reviewers.

3.5.4 Registered reports process. One of the authors – and the editors – thought that it was a problem to not have the final stage 1 manuscript agreed on before data collection began, because it left an area of doubt over whether the authors' corrections to the requested stage 1 changes were satisfactory to the editors. Correcting the data collection process ahead of time is one of the areas of the publication process that registered reports are intended to fix. However, our implementation of the process retained some doubt in the process (see section 4 for more details on this topic).

3.5.5 Replication outcomes. We briefly summarise the four papers here, along with a description of what was changed in order to replicate the study, and what the outcome was of the replication:

- Shindler et al. [67] set out to replicate a 2018 study by Zehra et al. [75] on how students learn dynamic programming and the misconceptions that students have. Shindler et al. increased the sample size compared to the original study, used multiple institutions, and collected data online due to the COVID-19 pandemic. The misconceptions found in their study matched those from the original study (e.g. not recognising the correct recurrence relation), and found more areas of difficulty (e.g.inappropriate use of a brute force solution, not defining a proper base case).
- Finke et al. [25] aimed to replicate and extend the findings of the validation part of the Computational Thinking test (CTt) 2017 study by Román-González et al. [62]. Finke et al. used a German translation of the original CTt and also collected data online due to the pandemic. Like the original study, they found that reasoning and spatial abilities contributed uniquely and significantly to performance on the CTt, but they also found that additional variance was explained by complex numerical abilities (i.e., algebraic skills). Finke et al. replicated the finding that boys outperform girls on the CTt, but in contrast to the original study, found that these differences were largely independent of the gender differences in cognitive skills.
- Hundhausen et al. [41] performed a replication of Buffardi [10] in order to examine objective measures of individual contributions to team projects. They expanded on the original study by including multiple institutions which varied in the software engineering courses offered. They largely replicated the original data sources and measures, but with some adjustments (e.g., to account for how each participating course assigned grades, collecting demographic data on student participants). Hundhausen et al. were able to replicate four of five significant findings from the Buffardi study; relative commit shares was found not to be a

predictor of peer contribution ratings. Additionally, they found more significant associations between subjective and objective metrics of individual contribution to team projects than in the original study.
- Fowler et al. [27] performed a study to replicate a slightly simplified hierarchy of reading, tracing and writing code skills from a study by Lopez et al. [45]. They 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 and cannot determine in what order should these skills be taught.

4 LESSONS LEARNED FROM THE SPECIAL ISSUE

While editing the special issue we encountered several challenges relating to replications and registered reports. In this section, we reflect on these challenges in turn, and give recommendations of how they could be avoided or handled better in future.

4.1 Replication Studies

We found that there was a large variety in how and in which sections in the paper the authors contrasted their own study to the original study being replicated. We ended up having to ask some of the author teams to change the organisation of their manuscript after they had completed the full paper. Reflecting on this, we would recommend that there are guidelines or templates in place so that authors can already make the differences and similarities between the studies clear from the first stage manuscript onward. This would also make explicit to what degree the authors are aiming for an exact, empirical or conceptual replication (see Table 1).

For this special issue, we suggested, but did not require, that the author teams contact the authors of the original paper (see subsection 3.1). Although evidence from the psychology community suggests replications are more successful when at least one author of the original publication is involved [48], it should be noted that there are two ways to view this result. The advantage of including the original authors is that they can help provide the exact details and materials for the replication (although ideally this would be present in the original publication). Peterson and Panofsky [59] suggest that this is particularly useful when "task uncertainty" is high, such as in chemistry and physics when using a newly invented piece of equipment or novel chemical procedure. The danger is that the original authors may bias the study to make sure it replicates their earlier work [69]. For this issue we did not allow the original authors to be listed as authors on the submitted replication, but this might have been too restrictive. A pragmatic way to deal with this in conjunction with registered reports could instead be that the original authors could only be involved in writing and gathering the materials for the stage 1 submission but not be involved in the subsequent data collection, scoring, analysis and interpretation of the data for the stage 2 submission.

Finally, one of the authors flagged that close replications might not be possible as specific technologies or programming languages are no longer in use. The implication is that in computer science education field, many replications will be of empirical or conceptual nature and need to be interpreted and compared to the original study accordingly. However, we think that is very important to conduct such replications, as it allows us to determine how our old knowledge transfers to the current technological context.

4.2 Registered reports

It is important for registered reports that the authors do their best to conduct the research as they described in the plan in the stage 1 report. When this is not possible (e.g., a change of data collection procedure because of the COVID-19 pandemic), the changes should be clearly documented and justified. We encountered four preventable difficulties with the relationship between the stage 1 report and the stage 2 publication:

(1) We approved the stage 1 manuscript conditional on a few reviewer and editor comments that the authors still had to implement before starting data collection (similar to the acceptance process at computer science education conferences). As described in subsection 3.5, this left authors with some ambiguity as to whether they had correctly addressed these concerns, which is exactly what registered reports are meant to prevent. In hindsight, we should have performed an additional editorial approval on the final version of the stage 1 manuscript.
(2) The stage 1 manuscripts are not externally visible and verifiable after publication of the stage 2 manuscript; they were only used for internal reviewing purposes. It would be useful for the computer science education community to have a database or platform (e.g., the Open Science Framework) to publish the stage 1 reports, before the authors conduct the study, thus providing verifiable proof for later readers as to what was planned. We encouraged authors to publish the stage 1 manuscripts on the Open Science Framework but none did, perhaps due to unfamiliarity with the whole process and not seeing any benefits of doing so. In hindsight, we should have mandated this step (as part of the final approval in the previous point) – and should have asked why this was in our evaluation survey.
(3) We found that authors rewrote or reorganised material in their manuscripts between stage 1 and stage 2, probably to improve readability of the final submission. However, this made the stage 2 review process more difficult and time-consuming, because the editors had to spend a lot of time checking whether the changes were purely cosmetic or whether the study details had been changed. In hindsight, we should have mandated that the stage 1 manuscript (or at minimum, the research design part) be carried over unmodified into the stage 2 submission.
(4) It is acceptable to make changes to the design and execution of the study after stage 1. But these should be done in a transparent way. We found that in the stage 2 submissions it was not always clear what changes had been made, if any. We recommended to authors during the stage 2 reviews that they add a section to explicitly describe changes made after the stage 1 review. In hindsight we should have made this recommendation from the outset, perhaps in a paper template, so that authors were clear about what was expected.

In short: stage 1 manuscripts should be approved in their final form, then published online (with an embargo if necessary – something the Open Science Framework already supports), then carried over unmodified into the stage 2 submission, with an additional section explicitly describing the changes made to the study since stage 1 (e.g., moving data collection online, or adjusting the analysis for an unforeseen confound).

Finding and retaining reviewers for follow up reviews is a known challenge for journal editors. In the case of registered reports, this is even more challenging as the time between reviewing the registered report and the final manuscript is even longer: potentially a year or more. Typically, a reviewer would be asked to review the first version of the registered report, then the revised registered report, and then again the final manuscript of the completed study. However, as the reviewers were involved in making suggestions on the improvement of study design, the review of the final report served more as a check that everything was done according to plan. As we were a team of three on our special issue, we decided to divide the reviewing of the final four manuscripts amongst ourselves (the editors) instead of asking the original reviewers again, with one primary and one secondary editor-reviewer for each paper. In general, this worked out well, but we acknowledge that it is important to have a good spread of knowledge on the editorial board to do this. This is why we recommend to also consider approaching subject matter reviewers, ideally the ones that reviewed the registered report, recruiting new ones, or using a pool of associate editors, who would also have to study the stage 1 submission before reviewing the submitted stage 2 manuscript.

Finally, we noticed that asking the author teams to conduct a power analysis is not straightforward. Power analyses are not yet very common in the computer science education community, and, for more complex analyses like structural equation modelling, the procedure of conducting them is quite challenging, potentially requiring a simulation in a statistical program like R instead of simply entering parameters in a power analysis tool like G*power. Moreover, even for a simple t-test entering the parameters in a power analysis tool is not straightforward. To do this, one needs an understanding of the parameters and knowledge about how to obtain and justify them. The most challenging parameter is the effect size [2]. For example, in order to calculate the required sample size for a specific power (typically .80), a researcher needs to make an educated guess of the effect size they expect to find. There are three ways to arrive at this: 1. Finding similar studies in the literature and see what kind of effect sizes they found (which was relevant for our special issue with replication studies, although unfortunately, the effect sizes were not always reported in the original studies), 2. conducting a pilot study (which comes with its own problems such as that effect sizes obtained with small samples are often not reliable) or 3. Use Cohen's recommendations (small effect: .25, medium effect: .5 large effect: .8) In the fields of psychology and cognitive science, the research team may ask a statistician to help them with or perform the power analysis for them and, depending on their further input to the paper, this person may join the author team. As power analyses are such an important instrument to determine what kind of, or in fact if any, conclusions can be drawn from the data, we recommend that the researchers in the computer science education community who conduct quantitative studies should expand their expertise in this area or find collaborators who are well versed in such statistical procedures. However, until this is common, it may be problematic to include it as a requirement for the stage 1 of registered reports, as we had wanted to do.

5 ARGUMENTS FOR (AND AGAINST) WIDER UPTAKE

Our special issue involved two simultaneous interventions: accepting only replications, and using the registered report publication process for the reviewing. In this section we will argue in turn why these two items should be taken up more widely, but also discuss the limitations and conditions, in particular in relation to non-positivist research.

5.1 Replications

The arguments for replications are widely known, but the problem is that they are still reasonably rare, with only 2% of CSER studies being replications [34]. The survey by Ahadi et al. [1] found that computer science education researchers' main concern about conducting replications was that they would not be published, and not be valued by peers and funders. One route to rectify this might be, like our special issue, to create special issues or particular conference tracks that only accept replications. This both increases confidence among researchers that they can publish such studies, and signals that they are considered valuable. An additional approach might be to create awards specifically targeted at replications. In CSER, SIGCSE has a new "test of time" award for papers that are at least ten years old that "recognizes an outstanding paper published in the SIGCSE community that has had meaningful impact on computing education practice and research." Perhaps it could also have an award related to conducting replications. Importantly, this award should not be contingent on the outcome of the replication, to avoid introducing new bad incentives (where the replicators would bias the replication in order to become eligible for the award).

5.2 Registered reports

Henderson [38] provides an overview of the reasons why researchers should move from the traditional way of submitting papers about finished studies to registered report procedures. In her paper she distinguishes between benefits for the research community and benefits for the researchers themselves. For the research community, she lists examples like ensuring that the papers are of high quality and report reliable results, the reduction of researcher bias, including p-hacking and HARKing, and the elimination of publication and outcome bias. Benefits for the researchers themselves include peer review at a point in time when it is most helpful, a guaranteed publication[1] and reduction of stress, because publication is not contingent on the outcomes of the study. This in turn, makes sure that researchers no longer feel pressured to dress up the results, which in turn benefits the general quality of publications. More generally, we would also like to introduce a "humanitarian" argument: at the moment, a lot of time and resources, both from

[1]This is especially important for PhD students and early career (pre-tenure) researchers, who may need a publication by a particular deadline for career progression.

researchers and participants are wasted when studies are rejected after being carried out. It is better for all involved if the rejection happens earlier, or can be avoided by correcting the design while it is still possible to do so.

Registered reports generally require individualised deadlines for completion of the stage 2 manuscript. This means they are naturally suited to journals where publication dates are flexible. However, the Mining Software Repositories (MSR) conference has recently introduced a model where the stage 1 manuscript is published at the MSR conference and the stage 2 manuscript is published in the Empirical Software Engineering journal. This kind of hybrid model may be a possibility in CSER, which features both conferences and journals as popular publication venues.

We explicitly do not believe that registered reports are suitable for all types of publication. Some data analysis projects are purely exploratory, some papers are purely retrospective and thus a registered report may not be applicable. One larger area where registered reports (and replications) may not be as applicable is non-positivist work, which we address in the next section.

5.3 Non-positivist work

Publications in CSER feature a wide variety of different research approaches, from quantitative (which tends to dominate [37, 65]) to qualitative. Although research is often presented as a dichotomous split into quantitative and qualitative *methods*, a more fundamental point is that these methods usually originate from different *philosophies* of science. Quantitative methods are typically used with a positivist or post-positivist perspective. Positivism assumes there is an objectively observable ground truth [14], while post-positivism softens the belief in an absolute truth [57]. Thus it is inherent in positivism that multiple observers (such as an original study, and a replication) should be able to observe similar results if their findings both reflect this ground truth [60]. To distinguish their observations from random chance, [post-]positivist scientists often use null-hypothesis significance testing, with its (in)famous $p \leq 0.05$ check. The analysis can usually be specified in advance of data collection, which is why the approach fits so well with the registered report concept[2].

Qualitative scientists usually adopt other philosophies of science, such as phenomenological, constructivist, or transformative approaches [18, 40]. These approaches differ from [post-]positivist views in various ways, for example by accepting that each person has a different "truth", acknowledging that all science is subjective, or centring social transformation. While these philosophies may be comparatively less prevalent in CSER, as Tenenberg [70] notes: "it is also important for the nascent interdiscipline of computing education to open the space of inquiry to include a broad range of theoretical perspectives. We need cognitivism, socioculturalism, and many more besides."

From these perspectives, replication may not feel like an applicable concept: investigations are usually very context-dependent. Furthermore, qualitative methods tend to be iterative, engaging in cycles of collecting, processing, exploring, and analysing data. It is not as clear here that the data analysis process can be meaningfully verified ahead of data collection taking place. In this paper we have primarily discussed questionable *quantitative [post-]positivist* research practices. Qualitative research tends to have a different set of questionable research practices, such as lack of transparency, and too much subjectivity [55]. Noble and Smith [55] suggest that "if qualitative methods are inherently different from quantitative methods in terms of philosophical positions and purpose, then alternative frameworks for establishing rigour are appropriate" – such as the criteria proposed by Tracy [71] or the strategies proposed by Shenton [66].

For our special issue, we explicitly invited qualitative as well as quantitative studies. However, we only received quantitative submissions, which may be a sign that qualitative researchers did not see how their methods would fit into this process. Replications and preregistration procedures might be less straightforward for qualitative than for quantitative studies, but procedures and templates do exist. For instance, Haven et al. [35] developed a template for preregistration of qualitative studies, and the cognitive science journal 'Cortex' accepts registered reports for qualitative studies. Chambers and Tzavella [12] discuss the need for developing more resources for qualitative registered reports.

With regard to replications, Talkad Sukumar and Metoyer [69] discuss how the interpretation of replication needs to be redefined for qualitative inquiry based on its nature and focus on interpretation. They suggest that of the types we show in Table 1, only conceptual replication may be applicable for qualitative research.

With regard to registered reports, we reject the idea that registered reports are inapplicable to qualitative work – but they may need adjusting to fit. The idea of registered reports is to evidence the research process and thus increase transparency. Lack of transparency is an issue for both quantitative and qualitative analysis, so registered reports should be able to aid in both. It may be that the structure needs to be slightly different for quantitative and qualitative work. Qualitative work, with its more iterative analysis, may benefit more from a stage 1 review (or stage 1.5 review) after data collection but before deep analysis. As Morse et al. [52] state in their paper on improving reliability and validity in qualitative research:

> "Regardless of the standard or criteria used to evaluate the goal of rigor, our problem remains the same: they are applied after the research is completed, and therefore are used to judge quality. Standards and criteria applied at the end of the study cannot direct the research as it is conducted, and thus cannot be used pro-actively to manage threats to reliability and validity." [52, pg .20]

This is precisely the problem that registered reports aim to solve; the main question is where during the work the additional review stage (or even stages) may be most appropriate and effective. We would welcome further work from qualitative non-positivist researchers to try out registered reports and evaluate whether and how they may be beneficial in CSER.

[2]There is potentially an argument to be made that registered reports are primarily fixing some of the issues introduced by statistical significance testing and its binary outcome that introduces the concept of positive/negative results, something which is not typically present in qualitative methods.

6 CONCLUSION

Career-based imperatives to "publish or perish", and a bias towards publishing positive results, can incentivise researchers to engage in questionable research practices in order to guarantee a publication. This includes manipulating elements of the study – such as the hypothesis or the data analysis – to ensure a positive result. This can lead to published research findings that are not an accurate reflection of reality, resulting in serious problems when stakeholders – such as teachers or national agencies – try to develop interventions or make policy decisions on this shaky basis. In this paper we have discussed two main defences against such practices: independently replicating the work of others to increase confidence in findings, and using registered reports – where research design is submitted and reviewed before data collection – to evidence the absence of such practices. We presented a case study of a journal special issue in the Computer Science Education journal where we only published replications, and only using a registered report process. We would inevitably make a few decisions differently in hindsight, but based on the author feedback and our own experience we strongly believe that replications should be further encouraged, and registered reports offer a good option for publication. We believe the current standard review process should be retained for research where registered reports may not be appropriate: registered reports are an alternative, not a replacement. As it stands, computer science education researchers estimate that only 30% of work is likely to replicate [1]; we believe that with structural changes to the publication process such as introducing registered reports, we can increase this number together.

ACKNOWLEDGMENTS

We are grateful for the support and encouragement of Jan Vahrenhold and Brian Dorn, the editors of Computer Science Education, who initiated this special issue and were patient with three novice editors. We thank the authors and reviewers of the papers in the special issue, who took a chance on registered reports, even though it was a new process at the journal. We thank Sue Sentance for some pointers about non-positivist approaches to research, and finally we are grateful to the ICER reviewers of this paper for their detailed suggestions, which helped to improve the paper.

REFERENCES

[1] 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* (Koli, Finland) *(Koli Calling '16)*. Association for Computing Machinery, New York, NY, USA, 2–11. https://doi.org/10.1145/2999541.2999554

[2] Marjan Bakker, Coosje L. S. Veldkamp, Olmo R. van den Akker, Marcel A. L. M. van Assen, Elise Crompvoets, How Hwee Ong, and Jelte M. Wicherts. 2020. Recommendations in pre-registrations and internal review board proposals promote formal power analyses but do not increase sample size. *PLOS ONE* 15, 7 (07 2020), 1–15. https://doi.org/10.1371/journal.pone.0236079

[3] Richard Border, Emma C. Johnson, Luke M. Evans, Andrew Smolen, Noah Berley, Patrick F. Sullivan, and Matthew C. Keller. 2019. No Support for Historical Candidate Gene or Candidate Gene-by-Interaction Hypotheses for Major Depression Across Multiple Large Samples. *American Journal of Psychiatry* 176, 5 (2019), 376–387. https://doi.org/10.1176/appi.ajp.2018.18070881 arXiv:https://doi.org/10.1176/appi.ajp.2018.18070881 PMID: 30845820.

[4] Richard Bornat. 2014. Camels and humps: a retraction. (2014). http://eis.sla.mdx.ac.uk/staffpages/r_bornat/papers/camel_hump_retraction.pdf

[5] Richard Bornat, Saeed Dehnadi, and Simon. 2008. Mental Models, Consistency and Programming Aptitude. In *Proceedings of the Tenth Conference on Australasian Computing Education - Volume 78* (Wollongong, NSW, Australia) *(ACE '08)*. Australian Computer Society, Inc., AUS, 53–61.

[6] Julia G. Bottesini, Mijke Rhemtulla, and Simine Vazire. 2022. What do participants think of our research practices? An examination of behavioural psychology participants' preferences. *Royal Society Open Science* 9, 4 (April 2022). https://doi.org/10.1098/rsos.200048

[7] Carter J. Boyd, Zachary L. Gentry, Kimberly D. Martin, and Soroush Rais-Bahrami. 2019. Factors Associated With the Highest and Lowest Cited Research Articles in Urology Journals. *Urology* 124 (2019), 23–27. https://doi.org/10.1016/j.urology.2018.11.034

[8] 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* (Espoo, Finland) *(ICER '18)*. Association for Computing Machinery, New York, NY, USA, 196–204. https://doi.org/10.1145/3230977.3230991

[9] Samuel V. Bruton, Mary Medlin, Mitch Brown, and Donald F. Sacco. 2020. Personal Motivations and Systemic Incentives: Scientists on Questionable Research Practices. *Science and Engineering Ethics* 26, 3 (01 Jun 2020), 1531–1547. https://doi.org/10.1007/s11948-020-00182-9

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

[11] Michael E. Caspersen, Kasper Dalgaard Larsen, and Jens Bennedsen. 2007. Mental Models and Programming Aptitude. *SIGCSE Bull.* 39, 3 (jun 2007), 206–210. https://doi.org/10.1145/1269900.1268845

[12] Christopher D. Chambers and Loukia Tzavella. 2022. The past, present and future of Registered Reports. *Nature Human Behaviour* 6, 1 (01 Jan 2022), 29–42. https://doi.org/10.1038/s41562-021-01193-7

[13] Christina S. Chhin, Katherine A. Taylor, and Wendy S. Wei. 2018. Supporting a Culture of Replication: An Examination of Education and Special Education Research Grants Funded by the Institute of Education Sciences. *Educational Researcher* 47, 9 (Dec. 2018), 594–605. https://doi.org/10.3102/0013189X18788047 Publisher: American Educational Research Association.

[14] Alexander M. Clark. 1998. The qualitative-quantitative debate: moving from positivism and confrontation to post-positivism and reconciliation. *Journal of Advanced Nursing* 27, 6 (1998), 1242–1249. https://doi.org/10.1046/j.1365-2648.1998.00651.x arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1046/j.1365-2648.1998.00651.x

[15] Andy Cockburn, Pierre Dragicevic, Lonni Besançon, and Carl Gutwin. 2020. Threats of a Replication Crisis in Empirical Computer Science. *Commun. ACM* 63, 8 (July 2020), 70–79. https://doi.org/10.1145/3360311

[16] Enrico Coiera, Elske Ammenwerth, Andrew Georgiou, and Farah Magrabi. 2018. Does health informatics have a replication crisis? *Journal of the American Medical Informatics Association* 25, 8 (04 2018), 963–968. https://doi.org/10.1093/jamia/ocy028 arXiv:https://academic.oup.com/jamia/article-pdf/25/8/963/34150203/ocy028.pdf

[17] Open Science Collaboration. 2015. Estimating the reproducibility of psychological science. *Science* 349, 6251 (2015), aac4716. https://doi.org/10.1126/science.aac4716 arXiv:https://www.science.org/doi/pdf/10.1126/science.aac4716

[18] John W. Creswell. 2008. *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches, 3rd Edition* (3rd edition ed.). SAGE Publications, Inc, Thousand Oaks, Calif.

[19] Jennifer Crocker and M. Lynne Cooper. 2011. Addressing Scientific Fraud. *Science* 334, 6060 (2011), 1182–1182. https://doi.org/10.1126/science.1216775 arXiv:https://www.science.org/doi/pdf/10.1126/science.1216775

[20] Joshua Cuevas. 2015. Is learning styles-based instruction effective? A comprehensive analysis of recent research on learning styles. *Theory and Research in Education* 13, 3 (2015), 308–333. https://doi.org/10.1177/1477878515606621 arXiv:https://doi.org/10.1177/1477878515606621

[21] Saeed Dehnadi and Richard Bornat. 2006. The camel has two humps (working title). (2006). http://eis.sla.mdx.ac.uk/research/PhDArea/saeed/paper1.pdf

[22] Brian D. Earp and David Trafimow. 2015. Replication, falsification, and the crisis of confidence in social psychology. *Frontiers in Psychology* 6 (2015). https://doi.org/10.3389/fpsyg.2015.00621

[23] Daniele Fanelli. 2009. How Many Scientists Fabricate and Falsify Research? A Systematic Review and Meta-Analysis of Survey Data. *PLOS ONE* 4, 5 (05 2009), 1–11. https://doi.org/10.1371/journal.pone.0005738

[24] Klaus Fiedler and Norbert Schwarz. 2016. Questionable Research Practices Revisited. *Social Psychological and Personality Science* 7, 1 (2016), 45–52. https://doi.org/10.1177/1948550615612150 arXiv:https://doi.org/10.1177/1948550615612150

[25] Sabrina Finke, Ferenc Kemény, Markus Sommer, Vesna Krnjic, Martin Arendasy, Wolfgang Slany, and Karin Landerl. 2022. Unravelling the Numerical and Spatial Underpinnings of Computational Thinking: a Pre-Registered Replication Study [in press]. *Computer Science Education* (2022).

[26] Center for Open Science. n.d.. *Registered Reports*. Retrieved 2022-03-16 from https://www.cos.io/initiatives/registered-reports

[27] Max Fowler, David Smith, Mohammed Hassan, Seth Poulsen, Matthew West, and Craig Zilles. 2022. Reevaluating the Relationship between Explaining, Tracing,

and Writing Skills in CS1 in a Replication Study [in press]. *Computer Science Education* (2022).

[28] Malte Friese and Julius Frankenbach. 2020. p-Hacking and publication bias interact to distort meta-analytic effect size estimates. *Psychological Methods* 25, 4 (2020), 456–471. https://doi.org/10.1037/met0000246

[29] Andrew Gelman and Eric Loken. 2014. The statistical crisis in science data-dependent analysis—a "garden of forking paths"—explains why many statistically significant comparisons don't hold up. *American scientist* 102, 6 (2014), 460.

[30] Stephen L. George. 2016. Research misconduct and data fraud in clinical trials: prevalence and causal factors. *International Journal of Clinical Oncology* 21, 1 (01 Feb 2016), 15–21. https://doi.org/10.1007/s10147-015-0887-3

[31] Stephen L. George and Marc Buyse. 2015. Data fraud in clinical trials. *Clinical investigation* 5, 2 (2015), 161–173. https://doi.org/10.4155/cli.14.116 25729561[pmid].

[32] Gowri Gopalakrishna, Gerben ter Riet, Gerko Vink, Ineke Stoop, Jelte M. Wicherts, and Lex M. Bouter. 2022. Prevalence of questionable research practices, research misconduct and their potential explanatory factors: A survey among academic researchers in The Netherlands. *PLOS ONE* 17, 2 (02 2022), 1–16. https://doi.org/10.1371/journal.pone.0263023

[33] David B. Grant, Gyöngyi Kovács, and Karen Spens. 2018. Questionable research practices in academia: antecedents and consequences. *European Business Review* 30, 2 (01 Jan 2018), 101–127. https://doi.org/10.1108/EBR-12-2016-0155

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

[35] Tamarinde L. Haven, Timothy M. Errington, Kristian Skrede Gleditsch, Leonie van Grootel, Alan M. Jacobs, Florian G. Kern, Rafael Piñeiro, Fernando Rosenblatt, and Lidwine B. Mokkink. 2020. Preregistering Qualitative Research: A Delphi Study. *International Journal of Qualitative Methods* 19 (2020), 1609406920976417. https://doi.org/10.1177/1609406920976417 arXiv:https://doi.org/10.1177/1609406920976417

[36] Megan L. Head, Luke Holman, Rob Lanfear, Andrew T. Kahn, and Michael D. Jennions. 2015. The Extent and Consequences of P-Hacking in Science. *PLOS Biology* 13, 3 (03 2015), 1–15. https://doi.org/10.1371/journal.pbio.1002106

[37] Sarah Heckman, Jeffrey C. Carver, Mark Sherriff, and Ahmed Al-zubidy. 2021. A Systematic Literature Review of Empiricism and Norms of Reporting in Computing Education Research Literature. *ACM Trans. Comput. Educ.* 22, 1, Article 3 (oct 2021), 46 pages. https://doi.org/10.1145/3470652

[38] Emma L Henderson. 2022. A guide to preregistration and Registered Reports. https://doi.org/10.31222/osf.io/x7aqr

[39] Christopher Lynnly Hovey, Lecia Barker, and Vaughan Nagy. 2019. Survey Results on Why CS Faculty Adopt New Teaching Practices. In *Proceedings of the 50th ACM Technical Symposium on Computer Science Education* (Minneapolis, MN, USA) *(SIGCSE '19)*. Association for Computing Machinery, New York, NY, USA, 483–489. https://doi.org/10.1145/3287324.3287420

[40] Kerry E Howell. 2012. *An introduction to the philosophy of methodology.* Sage.

[41] Christopher Hundhausen, Phillip Conrad, Adam Carter, and Olusola Adesope. 2022. Assessing Individual Contributions to Software Engineering Projects: A Replication Study [in press]. *Computer Science Education* (2022).

[42] Leslie K. John, George Loewenstein, and Drazen Prelec. 2012. Measuring the Prevalence of Questionable Research Practices With Incentives for Truth Telling. *Psychological Science* 23, 5 (2012), 524–532. https://doi.org/10.1177/0956797611430953 arXiv:https://doi.org/10.1177/0956797611430953 PMID: 22508865.

[43] Daniel Kahneman. 2011. *Thinking, fast and slow.* Macmillan.

[44] Anton Kühberger, Astrid Fritz, and Thomas Scherndl. 2014. Publication Bias in Psychology: A Diagnosis Based on the Correlation between Effect Size and Sample Size. *PLOS ONE* 9, 9 (09 2014), 1–8. https://doi.org/10.1371/journal.pone.0105825

[45] Mike Lopez, Jacqueline Whalley, Phil Robbins, and Raymond Lister. 2008. Relationships between Reading, Tracing and Writing Skills in Introductory Programming. In *Proceedings of the Fourth International Workshop on Computing Education Research* (Sydney, Australia) *(ICER '08)*. Association for Computing Machinery, New York, NY, USA, 101–112. https://doi.org/10.1145/1404520.1404531

[46] Hugues Lortie-Forgues and Matthew Inglis. 2019. Rigorous Large-Scale Educational RCTs Are Often Uninformative: Should We Be Concerned? *Educational Researcher* 48, 3 (2019), 158–166. https://doi.org/10.3102/0013189X19832850 arXiv:https://doi.org/10.3102/0013189X19832850

[47] Jonathan Lung, Jorge Aranda, Steve M. Easterbrook, and Gregory V. Wilson. 2008. On the Difficulty of Replicating Human Subjects Studies in Software Engineering. In *Proceedings of the 30th International Conference on Software Engineering* (Leipzig, Germany) *(ICSE '08)*. Association for Computing Machinery, New York, NY, USA, 191–200. https://doi.org/10.1145/1368088.1368115

[48] 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. https://doi.org/10.1177/1745691612460688 arXiv:https://doi.org/10.1177/1745691612460688 PMID: 26168110.

[49] Zacharias Maniadis, Fabio Tufano, and John A. List. 2017. To Replicate or Not to Replicate? Exploring Reproducibility in Economics through the Lens of a Model and a Pilot Study. *The Economic Journal* 127, 605 (10 2017), F209–F235. https://doi.org/10.1111/ecoj.12527 arXiv:https://academic.oup.com/ej/article-pdf/127/605/F209/26495669/ejf209.pdf

[50] 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 ACM Conference on International Computing Education Research* (Espoo, Finland) *(ICER '18)*. Association for Computing Machinery, New York, NY, USA, 133–141. https://doi.org/10.1145/3230977.3230984

[51] Alison McCook. 2017. *"I placed too much faith in underpowered studies:" Nobel Prize winner admits mistakes.* Retrieved 9th March, 2022 from https://retractionwatch.com/2017/02/20/placed-much-faith-underpowered-studies-nobel-prize-winner-admits-mistakes/

[52] Janice M. Morse, Michael Barrett, Maria Mayan, Karin Olson, and Jude Spiers. 2002. Verification Strategies for Establishing Reliability and Validity in Qualitative Research. *International Journal of Qualitative Methods* 1, 2 (2002), 13–22. https://doi.org/10.1177/160940690200100202 arXiv:https://doi.org/10.1177/160940690200100202

[53] Philip M. Newton and Mahallad Miah. 2017. Evidence-Based Higher Education – Is the Learning Styles 'Myth' Important? *Frontiers in Psychology* 8 (2017). https://doi.org/10.3389/fpsyg.2017.00444

[54] Pentti Nieminen, Gerta Rucker, Jouko Miettunen, James Carpenter, and Martin Schumacher. 2007. Statistically significant papers in psychiatry were cited more often than others. *Journal of Clinical Epidemiology* 60, 9 (2007), 939–946. https://doi.org/10.1016/j.jclinepi.2006.11.014

[55] Helen Noble and Joanna Smith. 2015. Issues of validity and reliability in qualitative research. *Evidence-Based Nursing* 18, 2 (2015), 34–35. https://doi.org/10.1136/eb-2015-102054 arXiv:https://ebn.bmj.com/content/18/2/34.full.pdf

[56] Ernest Hugh O'Boyle Jr, George Christopher Banks, and Erik Gonzalez-Mulé. 2017. The Chrysalis Effect: How Ugly Initial Results Metamorphosize Into Beautiful Articles. *Journal of Management* 43, 2 (2017), 376–399. https://doi.org/10.1177/0149206314527133 arXiv:https://doi.org/10.1177/0149206314527133

[57] Abdul Hameed Panhwar, Sanaullah Ansari, and Asif Ali Shah. 2017. Post-positivism: An effective paradigm for social and educational research. *International Research Journal of Arts & Humanities (IRJAH)* 45, 45 (2017).

[58] David Peterson. 2016. The Baby Factory: Difficult Research Objects, Disciplinary Standards, and the Production of Statistical Significance. *Socius* 2 (2016), 2378023115625071. https://doi.org/10.1177/2378023115625071 arXiv:https://doi.org/10.1177/2378023115625071

[59] David Peterson and Aaron Panofsky. 2021. Self-correction in science: The diagnostic and integrative motives for replication. *Social Studies of Science* 51, 4 (2021), 583–605. https://doi.org/10.1177/03063127211005551 arXiv:https://doi.org/10.1177/03063127211005551 PMID: 33764246.

[60] Karl Popper. 2002. *The logic of scientific discovery* (second ed.). Routledge.

[61] Felipe Romero. 2019. Philosophy of science and the replicability crisis. *Philosophy Compass* 14, 11 (2019), e12633. https://doi.org/10.1111/phc3.12633 arXiv:https://compass.onlinelibrary.wiley.com/doi/pdf/10.1111/phc3.12633 e12633 PHCO-1228.R1.

[62] Marcos Román-González, Juan-Carlos Pérez-González, and Carmen Jiménez-Fernández. 2017. Which cognitive abilities underlie computational thinking? Criterion validity of the Computational Thinking Test. *Computers in Human Behavior* 72 (2017), 678–691. https://doi.org/10.1016/j.chb.2016.08.047

[63] Anne M. Scheel, Mitchell R. M. J. Schijen, and Daniël Lakens. 2021. An Excess of Positive Results: Comparing the Standard Psychology Literature With Registered Reports. *Advances in Methods and Practices in Psychological Science* 4, 2 (2021), 25152459211007467. https://doi.org/10.1177/25152459211007467 arXiv:https://doi.org/10.1177/25152459211007467

[64] Marta Serra-Garcia and Uri Gneezy. 2021. Nonreplicable publications are cited more than replicable ones. *Science Advances* 7, 21 (2021), eabd1705. https://doi.org/10.1126/sciadv.abd1705 arXiv:https://www.science.org/doi/pdf/10.1126/sciadv.abd1705

[65] Judy Sheard, S. Simon, Margaret Hamilton, and Jan Lönnberg. 2009. Analysis of Research into the Teaching and Learning of Programming. In *Proceedings of the Fifth International Workshop on Computing Education Research Workshop* (Berkeley, CA, USA) *(ICER '09)*. Association for Computing Machinery, New York, NY, USA, 93–104. https://doi.org/10.1145/1584322.1584334

[66] Andrew K Shenton. 2004. Strategies for ensuring trustworthiness in qualitative research projects. *Education for information* 22, 2 (2004), 63–75.

[67] Michael Shindler, Natalia Pinpin, Mia Markovic, Frederick Reiber, Jee Hoon Kim, Giles Pierre Nunez Carlos, Mine Dogucu, Mark Hong, Michael Luu, Brian Anderson, Aaron Cote, Matthew Ferland, Palak Jain, Tyler LaBonte, Leena Mathur, Ryan Moreno, and Ryan Sakuma. 2022. Replication of "Student Misconceptions of Dynamic Programming" [in press]. *Computer Science Education* (2022).

[68] E. Soloway. 1986. Learning to Program = Learning to Construct Mechanisms and Explanations. *Commun. ACM* 29, 9 (sep 1986), 850–858. https://doi.org/10.1145/

6592.6594

[69] Poorna Talkad Sukumar and Ronald Metoyer. 2019. Replication and transparency of qualitative research from a constructivist perspective. (2019). https://doi.org/10.31219/osf.io/6efvp

[70] Josh Tenenberg. 2014. Asking Research Questions: Theoretical Presuppositions. *ACM Trans. Comput. Educ.* 14, 3, Article 16 (sep 2014), 8 pages. https://doi.org/10.1145/2644924

[71] Sarah J. Tracy. 2010. Qualitative Quality: Eight "Big-Tent" Criteria for Excellent Qualitative Research. *Qualitative Inquiry* 16, 10 (2010), 837–851. https://doi.org/10.1177/1077800410383121 arXiv:https://doi.org/10.1177/1077800410383121

[72] Rivka Tuval-Mashiach. 2021. Is replication relevant for qualitative research? *Qualitative Psychology* 8, 3 (2021), 365–377. https://doi.org/10.1037/qup0000217

[73] Dylan Wiliam. 2022. How should educational research respond to the replication "crisis" in the social sciences? Reflections on the papers in the Special Issue. *Educational Research and Evaluation* 0, 0 (2022), 1–7. https://doi.org/10.1080/13803611.2021.2022309 arXiv:https://doi.org/10.1080/13803611.2021.2022309

[74] Yu Xie, Kai Wang, and Yan Kong. 2021. Prevalence of Research Misconduct and Questionable Research Practices: A Systematic Review and Meta-Analysis. *Science and Engineering Ethics* 27, 4 (29 Jun 2021), 41. https://doi.org/10.1007/s11948-021-00314-9

[75] Shamama Zehra, Aishwarya Ramanathan, Larry Yueli Zhang, and Daniel Zingaro. 2018. Student Misconceptions of Dynamic Programming. In *Proceedings of the 49th ACM Technical Symposium on Computer Science Education* (Baltimore, Maryland, USA) *(SIGCSE '18)*. Association for Computing Machinery, New York, NY, USA, 556–561. https://doi.org/10.1145/3159450.3159528