How Do Computing Education Researchers Talk About Threats and Limitations?

Kate Sanders
Rhode Island College
Providence, RI, USA
ksanders@ric.edu

Jan Vahrenhold
University of Münster
Münster, Germany
jan.vahrenhold@uni-muenster.de

Robert McCartney
University of Connecticut
Storrs, CT, USA
robert.mccartney@uconn.edu

ABSTRACT

Background and Context: Empirical researchers have a long-standing tradition of explicitly discussing the threats to and limitations of their research. In the past twenty years, these discussions emerged as a standard component of empirical research papers in computing education research (CER) as well.

Objective: Our goal was to find out how the CER community talks about threats and limitations (which we refer to collectively as "challenges") and how they "respond" to these challenges.

Method: Our dataset included a total of 77 papers from four venues: two CER journals, *Computer Science Education* (CSEJ) and *ACM Transactions of Computing Education* (TOCE); one CER conference, *ACM Conference on International Computing Education Research* (ICER); and, for comparison, one mathematics-education journal, *The Journal for Research in Mathematics Education* (JRME). We analyzed the discussions of threats and limitations in these papers using deductive codes drawn from the literature, while also being open to new codes that emerged from the data. We took the papers on their own terms, so where qualitative and quantitative papers discuss challenges differently, we report on both.

Findings: We found that the majority of these papers—65 out of 77—did discuss challenges. Depending on the research methodology employed, authors reported challenges that we could map to either internal and external validity, construct validity, and statistical conclusions or to trustworthiness criteria. Most of the challenges related to study design, but some related to study implementation, analysis, and interpretation of the results. Almost none of the challenges we found were unique to CER, but were also found in JRME. Our contributions include describing the broad range of challenges and responses that are discussed and connecting them to concrete circumstances in CER. Finally, we find that threats are seen both as important information about and as challenges to the study being reported. The tension between these two perspectives is resolved by including both, in patterns of challenge and response.

Implications: Most immediately, we hope that this paper will broaden the perspective of those sitting down to write a threats and limitations section. It may also be of use at the study-design phase, since some of the challenges can be avoided with the benefit of hindsight. And finally, we hope to start a conversation about the challenges to our research: which can be mitigated, when (and how) a convincing argument can be made that an apparent challenge is not a problem, and which challenges might warrant serious revisions before submission (by the author) or rejection (by reviewers).

This work is licensed under a Creative Commons Attribution International 4.0 License.

ICER '23 V1, August 07–11, 2023, Chicago, IL, USA
© 2023 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-9976-0/23/08.
https://doi.org/10.1145/3568813.3600114

CCS CONCEPTS

• **Social and professional topics → Computing education**.

KEYWORDS

computing education, threats, limitations, validity, trustworthiness

ACM Reference Format:
Kate Sanders, Jan Vahrenhold, and Robert McCartney. 2023. How Do Computing Education Researchers Talk About Threats and Limitations?. In *Proceedings of the 2023 ACM Conference on International Computing Education Research V.1 (ICER '23 V1), August 07–11, 2023, Chicago, IL, USA.* ACM, New York, NY, USA, 16 pages. https://doi.org/10.1145/3568813.3600114

1 INTRODUCTION

Twenty years ago, it was hard to find a discussion of threats and limitations in a computing education paper. TOCE did not exist in its current form; ICER did not hold its first conference until 2005; LaTiCE and CompEd are even newer. A search of the SIGCSE Symposium, ITiCSE, and *Computer Science Education* for 2003[1] retrieved a total of two papers [39, 42] that discussed threats and limitations. As a community, we discussed reliable software systems, security threats, stereotype threats, and the validity of tokens, keys, software, addresses, XML documents, loop invariants, and Java expressions, but not the validity of our research in computing education. This was in sharp contrast to established practice in, e.g., medical research where discussing limitations of study designs and empirical result can have life-saving effects—see, e.g., [23, 63] for editorial comments on this issue. While one might argue that CER impacts people in a different way than medical research, research designs in any discipline can lead to biased, unjust, or non-reproducible results and thus it is of utmost importance to clearly indicate possible limitations and threats such that reviewers and readers can decide to which extent results can be trusted and possibly applied in different contexts. In consequence, discussing threats and limitations has become common in empirical computing education papers as well.

In this paper, we investigate what discussions of threats and limitations in the CER literature actually say. In doing so, we conduct a methodological review in which, according to Randolph "a content analysis approach is used to analyze the research practices reported in a body of academic articles [and which thus] differ[s]

[1] For the papers in the ACM DL, we used `AllField:(threat* OR limit* OR valid* OR reliab* OR trust*)` and the venue's DOI as the search term. This resulted in 98 papers which, together with all 17 papers published in *Computer Science Education* that year, we then inspected manually to exclude false positives that discussed, e.g., teaching interventions or curricula related to **reliab**le networks or security **threat**s.

from meta-analyses in that research practices, rather than research outcomes, are emphasized" [65, p. 1]. While not particularly widespread, methodological reviews have been conducted in Computing Education Research, e.g., on research methods [65], the use of visualizations [74], the use of inferential statistics [72], and the collection and use of demographics [51, 58].

Our analysis includes both quantitative papers and qualitative papers. As we discuss in more detail in Section 2, the two approaches have different underlying philosophies that are reflected in different terminology: internal and external validity on the one hand, for example, and trustworthiness, credibility, and transferability on the other. For short, we call any threat to any of these a "challenge".

In our dataset, we found various ways of naming sections that discuss such challenges. Acknowledging "Threats and limitations" as a widely-used term, we, for brevity, refer to the part of a paper discussing any such challenges (whether in a named section or not), as a "threats section".

We approach this investigation from two different angles, corresponding to two different perspectives on a threats section. The first perspective, based in the literature, is that a threats section's purpose is to guide the reader in interpreting, applying, or extending the study described in the paper. As Fincher and Petre note, "With so many vectors of bias and threats to validity, vigilance is a constant necessity. But so is honesty ... [so that] evidence is exposed to scrutiny, to test and possible falsification" [18, p.74].

The first perspective (honest reporting) is evident in our data, but so is a second one. The second view sees the threats section as a dialogue between the researchers and some hypothetical challenger, perhaps someone posing a question when the work is presented, perhaps a reviewer, perhaps a reader. Thus, challenges are not just described; frequently they are accompanied by one or more responses. We will follow these two perspectives—the threats section as information for the readers and as a defense of the study being presented—throughout the paper.

We break down our over-arching research question, "How do computing education researchers talk about threats and limitations?" into the following two research questions:

RQ1 What challenges are raised in the computing education research literature?
RQ2 What responses are made, and how do authors relate these to challenges?

2 BACKGROUND

The notion of "validity" is frequently associated with the interpretation and use of test scores as set forth in Kane's work [33, 34]. In consequence, much of the work dealing with validity has focused on positivist, quantitative approaches and theories, but as we will outline below, qualitative methodologists have developed notions that resemble some aspects of validity. For a more detailed account of the history of theories and stances relative to validity and validation, which is beyond the scope of this paper, we refer the reader to the introductory chapter of Taylor's book [76].

Traditionally, clinical research has emphasized methodological rigor, including dealing with aspects of validity. In fact, Hulley et al. [30] use the process of making inferences from the actual subjects in a clinical study to the intended study sample and—via the accessible population—to the target population as a running example to guide the reader through all aspects of designing, executing, and evaluating a clinical study. In their wording, "validity resembles accuracy" but also adds "a qualitative dimension to considering how well a measurement represents the phenomena of interest" [30, p. 38]. They concede that not every phenomenon may be assessed relative to a ground truth or "gold standard" as some phenomena may be abstract or subjective "such as pain or quality of life" [30, p. 39].

The classical distinction between internal and external validity dates back to Campbell and Stanley [8], who define "internal validity" as "the basic minimum without which any experiment is uninterpretable" [8, p. 5] and "external validity" to be the degree to which an observed effect can be generalized to other "populations, settings, treatment variables, and measurement variables" [8, p. 5].

In her textbook on validity and validation in quantitative work, Taylor presents a sequence of research designs each of which is geared towards mitigating a certain threat to internal validity [76, Ch. 2]. Using this progression, she demonstrates that it is hard, if not impossible, to address all conceivable threats to internal validity. Instead, authors are obliged to present their results in a way that allows the readers to assess to which extent the results are plausible, generalizable, or both. Hulley et al. [30] concede that even for well-funded large-scale studies, there is always a trade-off between the degree of internal and external validity of a study and its feasibility.

On the surface, quantitative and qualitative research are similar: both pose questions, gather data, and report findings. Their underlying philosophies are quite different, however. As summarized by Lincoln and Guba, quantitative researchers, taking a positivist stance, assume that there is "a single, tangible reality 'out there' " and that "inquiry can converge on that reality"; qualitative researchers, taking an interpretivist or naturalist point of view, assume that there are "multiple constructed realities." Quantitative researchers assume that "[T]he inquirer and the objective of inquiry are independent", while qualitative researchers assume that "[T]he inquirer and the "object" of inquiry interact and influence one another" [44, p. 37].

These differences play out in the way research is conducted. Differences relevant to our discussion of threats and limitations include, for example, qualitative researchers' preference for purposive sampling—"a sampling method that focuses on very specific characteristics of the units or individuals chosen" [2]—over random sampling and their assumption that "the extent to which ... findings may be applicable elsewhere depend on the *empirical* similarity of sending and receiving contexts" [44, pp. 40–42].

As a result, different criteria for trustworthiness have emerged in some lines of qualitative research. Lincoln and Guba align "credibility" with internal validity and "transferability"—the ability to apply qualitative results in a new situation—with external validity. A challenge to credibility can be countered by such factors as lengthy and detailed observation, use of multiple data sources, and member-checking (testing the conclusions with the participants from whom the data were originally gathered), among others. Transferability is ensured by "providing sufficient descriptive detail" of the setting, researchers, participants, and analysis, because "the burden of proof lies less with the original investigator than with the person seeking to make an application elsewhere. The original inquirer cannot know the sites to which transferability might be sought, but the appliers can and do." [44, p. 298].

In summary, validity in quantitative research is often concerned with claims based on available evidence [76], whereas qualitative research approaches validity with a focus on a rich, detailed description of the setting, the researchers, the participants, the data gathered, the meaning of the data to the participants, and the analysis [49]. Nonetheless, researchers taking both approaches share the aim of convincing the reader of the rigor with which research results have been obtained and of the results' consistency with established frameworks or—where applicable—theories.

3 RELATED WORK

Within computing education research, the importance of discussing threats and limitations is discussed in the two handbooks on Computing Education Research [18, 35] and in Randolph's doctoral research [65, 66]. Outside of computing education research, there are papers that, like this one, identify a dataset of empirical papers and analyze how threats to validity are reported in those papers. In the remainder of this section, we first discuss work from empirical software engineering and second, some papers from outside of computer science.

3.1 Empirical software engineering

Feldt and Magazinius [17] published an overview of threats in empirical software-engineering papers. The Feldt and Magazinius paper looks at full research papers, both qualitative and quantitative, from a single year. Its dataset is drawn from a single "state-of-the-art" venue that the authors "expected to have higher expectations, standards, and experience of empirical research" [17, p. 375].

Feldt and Magazinius deductively classify threats into seven categories drawn from the literature on both quantitative and qualitative research: internal validity, construct validity, conclusion validity, external validity/ transferability, credibility, dependability, and confirmability. Only one type of response is considered, "a specific choice or action used to increase validity by addressing a specific threat" (called a "mitigation strategy") [17, p. 374]. The mitigation strategies are classified according to the phase of the study they affect: design, data-collection, or analysis. In addition, "future work" is considered as a phase and used to classify those mitigations that are only mentioned as possible future strategies. Results include the total numbers of threats and mitigation strategies found, with averages per paper, but not mitigations per threat type. Finally, Feldt and Magazinius report that the standard terminology they used in classifying threats was rarely used by the authors in the dataset.

Sjøberg et al. [75] examine threats to validity (among other topics) in a more specific domain: quantitative controlled experiments in which individuals or teams performed one or more software engineering tasks. Like Feldt and Magazinius, they use deductive content analysis to determine which of the papers in their dataset discuss threats and then to classify those threats, but with different categories. Instead of Feldt and Magazinius's seven categories, they look only at two: internal and external validity. These in turn are broken down into intermediate-level categories (some of which might elsewhere be counted as "conclusion validity" or "construct validity"). Internal validity threats are classified as selection, history, maturation, regression, attrition, testing, or instrumentation, and external validity threats are classified as threats to subject, task, environment, and/or treatment.

Like Feldt and Magazinius, Sjøberg et al. consider responses that "reduced" or "eliminated" a threat, here referred to as "control actions". Unlike Feldt and Magazinius, Sjøberg et al. do not consider the phase of the study in which mitigations or threats occur. Overall, they conclude that "A major finding is that the reporting [of threats to validity] is vague and unsystematic. The community needs guidelines [. . .]." [75, p. 749].

Neto and Conte [56] present a secondary analysis of quantitative controlled studies, based on a dataset drawn from major software engineering journals. Like Feldt and Magazinius and Sjøberg et al., Neto and Conte use deductive content analysis to classify threats, in this case using Feldt and Magazinius's four quantitative categories: internal validity, external validity, construct validity, and statistical conclusion validity. Like the two earlier papers, this one reports counts of threats and "actions to address threats", but the actions are not characterized or associated with threats or study phase. Instead, the discussion focuses on a flowchart-like model that would capture how actions mitigating one threat can lead to another.

Lenarduzzi et al.'s paper [40], which examines threats due to participant selection, illustrates the potential value of focusing on a specific source of threats. Their dataset is very broad, being derived, not from selected journals and conferences, as in the earlier studies, but from searches of the ACM Digital Library, IEEEXplore Digital Library, Science Direct, Scopus, Google Scholar, CiteSeer library, Inspec, and SpringerLink. The analysis, on the other hand, focuses on two specific questions: how participants were selected and whether any threats were mentioned in relation to participant selection. Despite the use of convenience samples in at least 90 of the 118 papers (the others did not report how their participants were selected), only 50 of the 118 papers reported threats related to participant selection, and none suggested any mitigation strategies. To put this in context, and as a basis for recommendations, the paper draws on the literature on threats to convenience samples and possible mitigation strategies.

Ampatzoglou et al. [3] report on an analysis of threats to *secondary* studies—studies where the dataset consists of published papers—in empirical software engineering. Like Feldt and Magazinius, Sjøberg et al., and Neto and Conte, they restricted their dataset to major software-engineering venues. Unlike the earlier papers, however, its analysis is inductive. Based on that analysis, it reports concrete threats to this particular type of study, along with concrete mitigation actions associated with each threat, and develops a checklist for use by study designers, authors, and reviewers.

3.2 Papers from other non-CER disciplines

A number of studies from other disciplines have used journal articles as data in examining limitations.

Ioannidis [31] examined a set of 400 articles, 50 from each of eight highly-cited scientific journals, and checked to see whether they mentioned anything about limitations. His approach was to search for keywords in the text, using forms of the words *limitation, caveat, caution, shortcoming, drawback,* and *weakness*, and then verify whether the found word had anything to do with limitations of the current work, as opposed to other uses of the word or a reference to a limitation in previous work. He found that limitations of current work was rarely present. *Limitation/caveat/caution* only showed up in the context of current work in 67 (17%) of the papers; *Shortcoming* or *drawback* showed up in the context of current work in only 2 papers, and *weakness* only showed up once. It should be noted that if an author used the term "Threats to validity" it would not have been found.

Brutus et al. [7] looked at the limitations and future work reported by authors in four management journals; 1267 papers from the years 1982, 1987, 1992, 1997, 2002, and 2007. They extracted each limitation and reference to future work, then categorized them by limitation type: internal validity, external validity, construct validity, statistical conclusion validity, or issues with theory. They further categorized each paper by management subarea and methodological choices (Study setting, Design type, Temporal perspective, and Data analysis). They found that most articles (62.5%) mentioned at least one limitation, most articles (64.4%) had at least one direction for future work, and that both of these percentages increased over time. Internal and External validity were the most common limitations reported.

Wang et al. [79] examined 81 observational (cohort or case-control) studies with clinical outcomes in six of the most prestigious medical journals; their data items included the articles, their abstracts, plus accompanying journal editorials, journal press releases, and news stories referring to these articles. Their focus was on causality limitations: whether they were present, and whether they were accompanied by a disclaimer, "a statement that undermines or downplays the limitation" [79, p. 1571]. Their key findings were that the number of sources that reported a causality limitation (inherent in cohort and case-control studies) was quite low: 22% of the studies mentioned this limitation in either the paper, its abstract, the accompanying journal editorial, or a journal press release, and this limitation was only reported in 10% of the news stories covering these studies. Moreover, 45% of these reported causality limitations were accompanied by a disclaimer. The authors suggest that the danger of under-reporting these limitations is that readers think the results indicate causal relationships, "promoting health practices based on evidence of modest quality. Up to 50% of such practices prove ineffective when tested in randomized clinical trials" [79, p. 1572].

Price et al. [62] looked at a particular limitation, low survey response rate, by examining all articles in seven general health education journals from 1990-2002 that used mail surveys and reported their response rates. They found that the average response rates at the the different journals ranged from 59.3% to 71.6%, that the percentage of papers with response rates less than 50% (by journal) ranged from 9.9% to 26.4%. They argue that because of the potential non-response bias, no articles with less than a 60% return rate should be published.

A number of authors published papers that support the reporting of limitations. Ross and Zaidi [67] explain the importance of and goals behind describing limitations and why some authors might fail to include them. They also present a guide to reporting limitations, which states that "The presentation of limitations should describe the potential limitations, explain the implication of the limitations, provide possible alternative approaches, and describe steps taken to mitigate the limitations" [67, p. 261]. They provide examples of specific threats organized by phase of the study (study design, data collection, data analysis), and related to either internal or external validity. In addition, there have been editorials supporting the reporting of limitations in a number of disciplines, for example Puhan et al. [64] in medicine, Price and Murnan [63] in health education, and Greener [23] in learning technology as well as domain-specific guidelines for discussing threats and limitations in, e.g., health education [29, 45] and pediatric psychology [14].

4 METHODS

We focus on challenges that are explicitly identified as such by the authors. Unlike Lenarduzzi et al. [40], discussed in Section 3, we do not attempt to infer challenges from papers' descriptions of their studies. For example, in quantitative papers, the discussion of statistical methods, usually contained in a separate Methods section, could be considered as an implicit response to a threat to Taylor's "validity of statistical conclusions". Similarly, such a Methods section would argue for why instruments used for measuring a construct are actually well-suited to do so—thus addressing Taylor's "construct-related evidence for validity", and the "thick descriptions" of context and methods emphasized in qualitative papers address an implicit threat to credibility and trustworthiness. That said, the heart of our analysis is not compiling a complete list of the threats to each study, but the discussion of challenges and responses.

As researchers, we have been active in Computing Education Research for well over a decade each, working with both qualitative and quantitative methods. Having read many threats sections (and written some) gave us a broad perspective. At the same time, to minimize bias, we were careful to anonymize the segments extracted from the data (see below), and in coding and interpreting those extracts, we constantly reminded ourselves that the purpose of our work was not to re-review the papers, but to answer our research question, namely how computing education researchers talk about threats and limitations.

4.1 Data sources

Our dataset is a purposive sample, with the goal of identifying papers that were likely to address threats and limitations. Thus, the dataset is drawn from four research-oriented venues: *Computer Science Education* (CSEJ) and *ACM Transactions on Computing Education* (TOCE) represent the CER journals, the *ACM Conference on International Computing Education Research* (ICER), the CER conferences[2], and the *Journal on Research in Mathematics Education* (JRME), a top-ranked journal on mathematics education, is included for comparison.

[2] While the ACM Technical Symposium on Computer Science Education and the ACM Conference on Innovation and Technology in Computer Science Educational also list "Threats to Validity and Limitations" as part of their reviewing rubric for papers in the Computing Education Research track (see, https://sigcse2023.sigcse.org/track/sigcse-ts-2023-papers#Instructions-for-Reviewers, https://iticse.acm.org/2023/paper-review-process/), the papers in these venues are not classified by the track they were submitted to. Including SIGCSE TS or ITiCSE papers thus would have confounded the sample as papers from the Experience Reports and Tools tracks inadvertently might have been included or papers from the Computing Education Research tracks might have been overlooked.

The resulting dataset includes a total of 77 papers. For ICER, we included all 30 full-length research papers from 2021 (from the August 2021 conference, the most recent at the time we began our analysis). For a comparable time period, we examined the journals from September 1, 2020 to August 31, 2021 (excluding special issues). The rationale for excluding special issues was to prevent our sample from being biased towards methods (and threats) specific to the topic of such a special issue. As a result, we included all 15 research papers for JRME and all 16 research papers outside of special issues for CSEJ. To roughly equalize the size of the sample from TOCE with the other two journals, we took the 16 most recent papers outside special issues (which meant the contents of the last two issues before September 1, 2021). Papers from this dataset cited in our paper are marked in the references section by an asterisk in front of the first author's first name.

In each paper, we looked for a separate (sub)section with a relevant name, such as "Threats to validity", "Limitations", "Limitations and future work", or "Trustworthiness". For those papers without such a section, we searched the text of the papers on the terms "threat∗", "limit∗", "valid∗", "reliab∗" and "trust∗" and examined any occurrences of those terms closely in context.

For each paper, we copied the relevant discussions into a single text file. To minimize any bias that might occur if the authors and their affiliations were readily available during analysis, each text file was named using the corresponding paper's DOI.

4.2 Analysis

We used MAXQDA to analyze and code the relevant discussions. Our basic unit of analysis was a single sentence. The end of a unit was indicated by the usual end-of-sentence punctuation, such as periods, question marks, and exclamation points. In addition, as we found that colons were sometimes followed by complete sentences and semi-colons by incomplete ones, we decided to insert a line (unit) break after each such punctuation mark that was followed by a complete sentence, unless the semicolon separated numbered items in a list (in this case, we inserted a break regardless of the grammar).

After some exploratory analysis, we converged on two top-level codes: "Challenge" and "Response". The code "Challenge" covers anything that might affect how the results presented in the paper should be interpreted and applied (whether a limitation, a threat, or a discussion of the trustworthiness of qualitative results). Since we are primarily interested in how CER researchers *talk* about threats and limitations, an item is not counted as a challenge unless the authors explicitly identify it as such. For example, suppose the authors state, "The data were analyzed by a single coder" in their Methods section. If the authors did not describe it as a challenge, we did not include it.

The code "Response" covers any answer to a challenge. Responses to the above challenge might include, for example, "We provided the coders with extensive training", or "Each piece of data was analyzed by two different coders and the inter-rater-reliability was x, which is considered very good." Again, we were only interested in how the authors talked about their response, not its quality.

Using the ICER papers as a pilot, we then coded each of the units of analysis using the challenge and response top-level codes, with all three researchers coding the papers individually. Aiming at "illustrating points of tension and [. . .] ultimately a stronger codebook" [50, Sec. 5.2.3], we did two full comparisons (researcher A to researcher B, then the merged file to researcher C), resolving differences through discussion in which all three researchers participated; as a result, we arrived at full agreement on all top-level codes.

Next, we returned to an exploratory analysis, examining the challenges to see if they fell into identifiable categories. For this task, we used a combination of deductive and inductive analysis. We looked at challenges identified in the literature, in particular Taylor [76] for quantitative papers and Lincoln and Guba [44] and Åkerlind [85] for qualitative work; we also noted challenges that emerged inductively from the ICER data. The response codes were developed inductively. To comply with Åkerlind's defensibility criterion, namely that each researcher should be able "to argue persuasively for the particular interpretation that they have proposed" [85, p. 330], each researcher first individually coded all of the ICER data. Then, codes were resolved through discussion; this resulted in full agreement on all lower-level codes.

Agreement on the ICER data gave the researchers "capacity to code more data" [50, Sec. 5.2.1]. To process the remaining data, the researchers each took a different one of the three journals and applied the codes developed for ICER, discussing any questions that arose with the group. Next, as a cross-check for our codebook and to strengthen defensibility, the researchers divided up the codes that had been identified ten or more times. Each researcher then took one or more of these codes and examined them across all four venues, to ensure that they had been applied consistently. Any questions were raised with the group and resolved through discussion. As a last step, the researchers as a group examined all remaining codes, i.e., those occurring fewer than ten times, making sure there were no additional issues. Thus, nearly all of the codes in the three journals were examined by at least two of the three researchers and we reached full agreement on all codes.

In addition to the codes used to indicate the type of challenge, we also used codes to indicate the phase of the research from which—according to the author's account—the challenge arose. We developed a code system that consisted of four codes: "Experimental design", "Experimental execution", "Data analysis", and "Interpretation" (comparable to the classification of mitigation actions used by Sjøberg et al. [75]). This code system does not judge the appropriateness of study design, research methodology used, or the type of data collected to answer a research question unless the authors mentioned this themselves and it thus became relevant for our research question.

Finally, in order to reason about the relationship between challenges and responses, we identified larger "challenge-response units", each of which contains the set of segments describing a single challenge and the responses to that challenge (if any). Figure 1 visualizes the challenge-response units in a single threats section. The units of analysis ("segments") are shown from left to right. For each segment we list all codes grouped by (sub)category from top to bottom. In the top lane, we indicate the study design. For a "challenge" segment, the next four lines ("When") are used

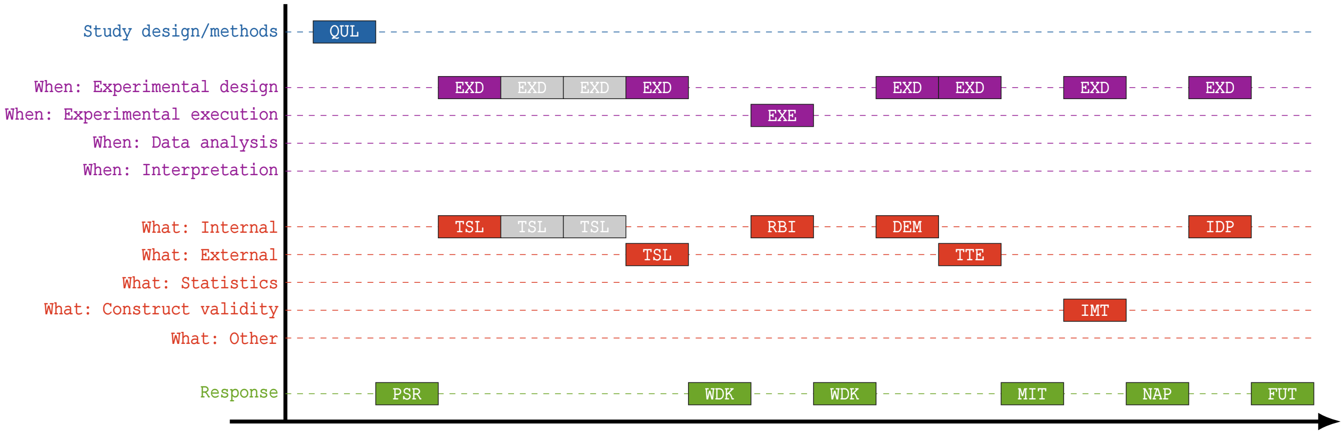

Figure 1: Visualization of challenges and responses using codes and shadow codes for one paper [11].

to indicate at which point of the process a challenge arose and the following lines ("What") are used to indicate the type of challenge according to the classification discussed in Section 4.

For a top-level classification, we used Taylor's four categories "Internal validity", "External validity" (both of which can be reinterpreted in a qualitative context), "Validity of statistical conclusions", and "Construct-related evidence for validity". To maintain an open perspective for challenges related to specific research methods possibly not covered by the other categories, we added a fifth ("Other") category. For responses, we use a separate lane; we refer the reader to Section 5.3 for a detailed discussion of the codes in this category.

As shown in Figure 1, this is a qualitative paper (indicated by the first code on the left, QUL). The authors start their discussion with a general disclaimer, "While we think this study sheds light on the potential for rubrics to structure and attune teachers in their evaluation of curricular materials, it is not without its limitations." [11]. Based upon the authors' choice of the words "sheds light", we coded this as a "Preliminary/suggestive results" response (PSR).

The next segments refer to a "Treatment-selection interaction" (TSL), an interaction between the selection of the participants and the treatment that might affect the internal validity. As the second and third segment in this group of segments are coded with the same three-letter code but rendered in gray, i.e., as a "shadow code", we see that these continue to describe the same challenge. This challenge has no associated response.

The next segment refers to a second "Treatment-selection interaction" that may or may not ("We don't know" response; WDK) affect external validity; the same response was given to a "Researcher bias" (RBI) challenge to internal validity. We then see a "Treatment-testing interaction" (TTE) challenge to external validity with the "Mitigated" (MIT); an "Inadequate measurement of target construct" (IMT) challenge to construct validity with the "Not a problem" (NAP); and an "Instrument design/presentation" (IDT) challenge to internal validity with the "Future work" (FUT). Note that, with the exception of "Researcher bias", all challenges were coded as arising from the study design.

5 RESULTS

The first result is that, as a community, we do talk about threats and limitations. As shown in Table 1, of the 77 papers in our dataset, 65 contained discussions of threats and limitations. These 65 papers include the large majority not only of the research papers in our dataset, but also of the papers in all four venues individually.

Moreover, the threats we are talking about, in general, are not unique to CER. Of the types of threat that occur in more than one paper in the dataset, all the threats in JRME were also found in one or more of the CER venues, and all but five of the challenges found in the CER venues were found in JRME.

In the remainder of this section, we first give an overview of the dataset, and then discuss the answers to each of our research questions in turn.

5.1 Characterizing the dataset

Of the 77 papers in our dataset, 10 were excluded for the reasons discussed in Section 4: they did not have a section or subsection with a relevant name, and a keyword search, combined with a reading of the paper, did not find any relevant discussion elsewhere in the paper. After we began analysis, two additional papers were excluded because their discussions focused on strengths only, without mentioning any challenges.

The discussions (where they were present) ranged from 2 to 72 coded segments per paper (with each segment being basically a sentence, as described in Section 4). There were only three documents with over 36 coded segments, the average number of coded segments in documents with a discussion was 15.35, the median number was 12. Note that there is no one-to-one mapping of codes to units of analysis. Rather, it depends on the paper's writing style. Two or more units of analysis are sometimes used to describe a single challenge or response; alternatively, a single unit may include more than one challenge or response, or a single challenge-response pair. In consequence, these counts are included only to approximately characterize the dataset.

The number of distinct challenges discussed in a single paper ranged from 1 to 14 with an average of 4.8 and a median of 4 across all papers. Distinct challenges are those raised by the paper that may

Table 1: The number of papers in our dataset from each venue, and how many of each contain discussions of threats and limitations. Named subsections and discussion elsewhere (that is, outside of a named (sub)section) are reported separately, because discussions outside a named section were more challenging to identify.

Venue	Year (conference) / Issues (journals)	Number of papers	Discussion of threats	In named (sub)section	Elsewhere in paper	No discussion
ICER	2021	30	26 (86%)	24	2	4
CSEJ	30(2), 30(4), 31(1), 31(3)	16	14 (88%)	13	1	2
TOCE	21(1), 21(3)	16	12 (75%)	10	2	4
All CER		*62*	*52 (84%)*	*47*	*5*	*10*
JRME	51(5), 52(1)–52(4)	15	13 (87%)	6	7	2
All		77	65 (84%)	52	13	12

be of the same type, but do not involve the same concrete problem. For example, two "Limited setting" challenges, one because the study was done at a single university, and a second, because it was done in a particular semester, would be considered distinct.

5.2 RQ1: The challenges

In this section, we present the challenges that were raised—that is, the challenges that were discussed in our dataset; we focus on challenges mentioned at least twice. These challenges were *not* necessarily flaws in the papers that raised them; the majority of the challenges were associated with some sort of response.

The section is organized according to concrete issues that might be encountered in a study. For convenience, these issues are grouped into the following (sometimes overlapping) categories: challenges related to the study's setting (Section 5.2.1), the study's participants (Section 5.2.2), researchers (Section 5.2.3), tasks and materials (Section 5.2.4), implementation (Section 5.2.5), data (Section 5.2.6), statistical validity (Section 5.2.7), construct validity (Section 5.2.8), limitations inherent in the type of study (Section 5.2.9), and specifics of qualitative analysis (Section 5.2.10). Each concrete issue is followed by a classification (generally a formal term from the literature, but sometimes derived inductively) and then by examples from our dataset.

5.2.1 Setting. We found three challenges related to a study's setting which, depending on the context, might affect internal or external validity.

Were the data gathered in some limited time or place or from a limited population? ("Limited setting", derived inductively from our data.) (28 occurrences).

This type of challenge, a threat to external validity, is discussed frequently in our dataset. Examples include—aside from the common "just one university"—a single school district, the United States, four countries in Europe, entering students, third-year students, and curriculum documents written for United States institutions, among others. The challenge is not related to any particular feature of the setting, simply to the fact that it is limited.

Is the study affected by some feature of its setting? ("Treatment-setting interaction" [76]) (19 occurrences).

This challenge can be either external or internal. Internal "Treatment-setting interaction" challenges include a situation where the participants might have put less effort into the intervention because it was a small part of a demanding project [83]; and a study in which the current participants were young children attempting to read analog clocks in an experimental setting and might have performed differently in a classroom or at home [15].

External "Treatment-setting interaction" challenges raise the possibility that some feature of the current setting makes it difficult to generalize. One clear example is the difference between a university setting and industry; one study of pair programming notes carefully that "we make no claims about the effectiveness of pair programming in industry settings, where both the implementation and the goals of pair programming vary from its use as a pedagogical tool in coursework" [6]. Other examples include the difficulty of generalizing from CS1 to other computer-science courses [82]; and from a mathematics teacher whose school administration was particularly supportive to teachers in other environments [68].

The key difference between "Limited setting" and "Treatment-setting interaction" challenges is that for a "Treatment-setting interaction" challenge, there must be some mention of an interaction between the setting and the study. For example, one paper raised both types of challenge. It noted that "all participating universities were located in the United States" ("Limited setting") and also that, specifically, they were research universities, and "one might expect that the findings would look very different in these other contexts (e.g., faculty interactions may play a stronger role at liberal arts colleges)" (a "Treatment-setting interaction") [5].

Was the study done during a major disruption, such as a hurricane? ("Treatment-history interaction" [76]) (8 occurrences).

In our data, the "major disruption" was the COVID-19 pandemic, which was discussed as a threat in several of the ICER papers [4, 10, 12, 37, 38, 53, 83, 84].

5.2.2 Participants. Several of the challenges we found are closely tied to the use of human participants. Three of these relate to the size or composition of the sample. They include:

Were the participants volunteers? ("Volunteer bias" [2]) (5 occurrences). The presence of volunteers raises the possibility that the sample might be unrepresentative of the larger population. Five papers in our dataset raised this issue, always in connection with a study involving student participants [19, 28, 46, 47, 84].

Are there enough participants or survey respondents? ("Small participation/response rate") (15 occurrences). This challenge focuses on the number of participants or the number who returned surveys, rather than the fact that they are volunteers (although that issue is also present).

Is the sample unrepresentative of the target population, or are the treatment and control groups different in ways that might affect the results? ("Treatment-selection interaction" [76]) (12 occurrences). Examples from our dataset include students who had prior experience with part of the material being tested [80], and control and experimental groups that were five years apart [26].

The remaining two challenges in this group relate to the quality of human-participant data:

Could the participants be adjusting what they say or do in response to what they think the researcher is looking for? ("Demand characteristics", sometimes called "sociability bias" [76]) (13 occurrences). Examples in our dataset include situations where the researchers [53] or the creators of the tool being evaluated [28] are present during interviews, or where the creator of the tool was the students' instructor [11]. In a study of student usage of GitHub, concern was expressed about students' "excessive committing to give the impression of a higher activity" in log data [78]. Another study noted that this challenge is "heightened for 'sensitive topics' such as gender issues" [10]. In the context of coursework, students may be influenced by clear assessment criteria (in the form of a checklist) [25] or by detailed instructor feedback [83].

Does the study rely on participants' descriptions of their feelings, attitudes, or behavior? ("Self-reported data" [2]) (6 occurrences). For example, in our dataset, this challenge is raised related to students reporting on their own skills [78], leadership confidence [5], level of stress [71], and approach to solving math problems [48], as well as faculty reporting their experiences running service programs during the COVID-19 pandemic [4].

5.2.3 *Researchers.*

There is one general category for threats attributable to the researchers themselves.

Do the researchers have any preconceptions or expectations that might affect the design, execution, or interpretation of the study? ("Researcher bias" [2]) (15 occurrences). We include under this "Rater bias", which occurs during analysis, concerns about the number or qualification of raters, and general discussions of the researchers' bias that might have an effect throughout the study.

Several possible sources of this challenge were suggested in our data. First, researchers might be biased by holding additional roles, such as being instructors in the class from which data are being gathered [11, 78, 83]. Second, researchers might be biased by additional information about the participants, for example how they performed on another portion of the dataset (student grades vs. student activities gathered from a log file [78]) or how they write [83]. Third, knowing the purpose of the study can bias those coding the data [26]. Fourth, using a deductive coding scheme can cause researchers to miss something that does not fit into one of the predetermined categories [81]. Finally, qualitative researchers assume that all researchers bring some biases to their research, so there is an emphasis on identifying and disclosing those biases. See, e.g., [27, 43].

5.2.4 *Tasks and materials.*

In our data, many references were made to potential problems with the materials or tasks given to participants that might affect the internal validity of a study. For these purposes, a "task" is something the participant is to do, and "materials" are something the participants are given to read or watch.

Were there problems with the materials or tasks given to the participants that might have affected the results of the study? ("Instrument design/presentation", inductively derived from our data) (39 occurrences).

Examples of task problems include, for example, "[T]he paper-and-pencil medium used during the interviews perhaps facilitated some syntax errors that might not have appeared while working on the computer." [52], and "[T]he situation differs from their regular debugging during coursework in that they [. . .] did not write the code themselves" [16]. An example of a materials problem is "Even though we tested the pictures in the pilot, it should be noted that the pictures we choose might result in different associations than the one we were aiming at." [12]. Sometimes authors combine aspects of both tasks and materials, as in "For example: most participants will have been unfamiliar with this type of task, and the video may have insufficiently prepared them for it" [84].

5.2.5 *Implementing the study design.*

There are potential issues with implementation when there are multiple people involved in running a study or when the participants (for whatever reason) may not have followed the instructions. We group these under a single category.

Were there multiple instructors or interviewers? Could the participants have failed to follow the instructions? ("Unreliable implementation" [76]) (27 occurrences).

In our dataset, for example, concerns were raised about inconsistency between interviews [1, 20], courses taught by multiple instructors [37, 41, 59], students possibly not following instructions [84] or cheating [54], and changes in implementation over time [26, 80].

5.2.6 *Data.*

We identified four challenges that are raised by various aspects of the data. They vary depending on whether the problem is with how the data were collected, how many items there are, what's missing, or the relationship between the sample data and the population. The challenges include:

Is the sample size too small? ("Small sample" [76]) (21 occurrences, including the "Small participation/response rate" challenge, discussed above in Section 5.2.2).

The "Small sample" challenge is found in situations that involve human participants, as described above, and also where recruiting human participants is not an issue. Examples in our dataset include cases where consent was not required for student data ([21, 78]), where the participant had volunteered, but the sample "only included four or five observations for each of the participant 's computing and collaborative behaviors" [32], and where the data were drawn from textbooks [13]).

Is there potentially useful data that was not collected, was not provided by the participants, or was collected but not analyzed? ("Incomplete data", derived inductively from our data) (26 occurrences).

In our dataset, sometimes part of the data was impossible to get, such as background information on participants (due to anonymity requirements) [61, 73]. Other issues mentioned included lack of data about whether students retained what they had learned [55], log data where student programs were included only when compiled [38], and papers possibly missing from a literature review [6]. Information not provided by participants included homeworks not turned in [21], post-tests not completed [6], and missing responses to survey questions [36]. Finally, sometimes not all the data available were included in the analysis [83].

Does the sample fail to match the target population in one or more characteristics relevant to the study? ("Nonrepresentative sample" [76]) (49 occurrences, including "Limited setting" (Section 5.2.1) and "Volunteer bias" (Section 5.2.2)).

As noted earlier, having volunteer participants always raises a possible "Nonrepresentative sample" challenge. Other examples in our dataset that raised this challenge involved samples with too few members of underrepresented groups [69, 83], too many poor and minority students [70], only grades 1–3 (approximately 6–8 years old) [60], only post-baccalaureate students [41], and students visiting a science museum [12]. One study raised this in relation to the choice of high-school students as proxies for future university students [10].

One paper described the difference between nonrepresentative and nonrandom samples particularly clearly, identifying a situation that is becoming more common as Institutional Review Boards add requirements for consent to what used to be exempt as "naturally occurring data". The original design for the study involved randomly selecting 50% of a large class to participate. The local Institutional Review Board required that the students be asked for consent, however. As a result, the study had a randomly selected sample (since the invitations were sent out at random), but only 40% of those invited, accepted. Upon analysis, it was discovered that the students who volunteered, were, on average, stronger students than those who did not, making the sample nonrepresentative [47].

5.2.7 Validity of statistical conclusions. Several threats to the validity of statistical conclusions were raised in only one or two threats discussions each, for example nonrandom samples [47], lack of a control group [80], nonrandom assignment to groups [26], the handling of missing responses to survey data [36], test assumptions not met [36], and failure to correct for multiple tests [77]. Small samples and nonrepresentative samples were discussed above in Section 5.2.6. Only one additional threat to statistical validity was discussed more than twice:

Is there a variable that was not taken into account that might have affected the study results? ("Omitted variable bias" [76]) (15 occurrences).

Papers that raise this threat usually explicitly mention "possible confounding variables" or missing "factors", "aspects", "mediating variables" or "potentially relevant constructs". The missing variables are fairly specific to the research question. For example, an investigation of computational thinking skills in primary school children used a test of cognitive ability, but noted that other missing variables might "include curiosity, creative talent, work habits, and study skills" [70]. A study that examined sense of belonging in university computer science students found that "both prior experience and incoming sense of belonging separately contributed to the model", suggested "that sense of belonging is an independent measure of student outcomes", but noted that sense of belonging might be related to peer networks [37].

5.2.8 Construct validity. "Construct validity" (also referred to as "Construct-related evidence for validity" [76]) is the fourth of the standard high-level threat categories, along with internal validity, external validity, and statistical conclusion validity. A "construct" is a "set of related behaviors and/or cognitive processes that are grouped together and named" [76, p. 4]. Specifically, constructs are the phenomena we are investigating: generally invisible phenomena, such as understanding recursion, motivation to learn programming, confidence, and so on. Within this high-level category, we found one type of challenge in our dataset:

Are the measures used adequate? ("Inadequate measure of target construct" [76]) (30 occurrences).

This challenge was raised in our dataset, for example, in connection with "using students' grades as a proxy for judging the quality of their work" [78], the difficulty of designing a perfectly accurate search for a literature review [24], and a survey design in which "small number of items (seven items for each dimension) might not fully cover the range of the targeted constructs" [36].

5.2.9 High-level study design. Some aspects of the results of a study follow from the nature of the study: a study is qualitative or quantitative; data are gathered in the field or in a more controlled lab setting; the findings include correlations or causation, and so forth. Each of these choices has trade-offs, providing some benefits and imposing some costs. The limitations imposed by such choices were mentioned frequently enough in our dataset that we added a category to capture them:

Were important limitations imposed by the nature of the study design? ("Inherent limitations", inductively derived from our data) (19 occurrences).

Examples in our dataset include field studies [16, 41], studies in a lab setting [15], qualitative studies [57], quantitative studies [6, 61], and correlational studies [9].

5.2.10 *Qualitative analysis.* The criteria for trustworthiness in qualitative work discussed above in Section 2 suggest a different perspective on challenges. Researcher bias is a threat to qualitative work, as it is to work in the positivist tradition. Qualitative papers are more likely, however, to include a statement of the researchers' biases in response to this threat, e.g. [27, 43]. In addition, we found one paper whose threats section contained a particularly careful discussion of steps taken to mitigate those biases [73].

An inadequate sample is also a threat to both qualitative and quantitative work. The definition of "adequate" is different, however. The goal in qualitative work may be to have a sample that represents the range of possible perspectives on the question being investigated, rather than the population, so the challenge would be a sample that didn't include the necessary range of perspectives. Barker et al. raise this challenge, reporting that only faculty perspectives had been included when investigating service-learning partnerships with community organizations [4].

Sentance and Waite respond to an implicit threat to credibility (the analogue to "internal validity", as discussed in Section 2 above), by "using established research methods, by ensuring we were completely familiar with the context of the participating teachers, developing a relationship with the participants that would support integrity and honesty in their reports, by focusing on their lived experiences rather than opinions, and by using field notes memos within the QDA software as a reflective commentary" [73].

5.3 RQ2: Responses and challenge-response dialogues

While the papers in our dataset identify a large variety of challenges, most of the challenges raised (about three quarters) are addressed by some response. We developed the response codes inductively, as described in Section 4, finding a total of 17 different codes.

The more common codes describe most of the responses; the seven most common codes—those occurring ten or more times—cover over 85% of the responses. These codes are:

"Future work" (85 occurrences) This response suggests that further study could be used to address challenges. It is the most common response to challenges in our dataset and can be used as an overall response to multiple challenges in a study. It was the most common response for "Instrument design/presentation", "Demand characteristics", and "Researcher bias".

"Mitigated" (66 occurrences) This response describes what was done to reduce the effects of a given challenge. It was the most common response for"Instrument design/presentation", "Demand characteristics", and "Researcher bias".

"Not a problem" (54 occurrences) This code response states that the challenge is not an issue for their study results (often, but not always, with evidentiary support). It is the most common response to "Unreliable implementation", and often used for "Incomplete data" and "Instrument design/presentation".

"Limit inferences" (23 occurrences) This response clearly delineates the extent of the findings and the degree to which they may be generalized. It is most commonly a response to "Limited setting" or "Treatment-setting interaction".

"We don't know" (19 occurrences) This response is an indication that the author cannot say whether a particular challenge affected the paper's results. It was the most common response (tied with "Future work") to "Treatment-setting interaction", and fairly common for "Inherent limitations".

"Preliminary/suggestive results" (16 occurrences) This response is generally an admission that some challenge(s) likely threatened or limited the results of the study, but the study results still have value. It is most commonly used as a response to "Small participation/response rate", "Omitted variable bias", and "Limited Setting" challenges, and often includes terms like "preliminary", "tentative", and "suggestive".

"Trade-off" (14 occurrences) This response is a justification of some design or action that achieves a balance between two desirable but incompatible features. It is most commonly a response to "Treatment-setting interaction".

In the remainder of this section, we return to the notion of dialogue mentioned in Section 1, the repeated challenge-response pairs that we found in our data. We give examples of these challenge-response dialogues, showing each of the most frequent responses and each of the most frequent challenges in at least one dialogue.

5.3.1 *Setting.* We start with three different challenge-response pairs concerning the "Limited setting" challenge.

- *Limited setting? Future work.* This dialogue proposes that future work could determine whether findings can be generalized: " This finding may be a reflection of the pedagogy employed in computer science coursework at this institution, and potentially, other institutions. [. . .] this approach to documenting conceptions of learning computer science needs to be replicated to provide more conclusive evidence" [81].
- *Limited setting? Limit inferences.* This dialogue strengthens the simple "Future work" response, "delineating the extent of the findings", in terms of the definition of "Limit inferences" above, and arguing that the study in question has not solved the whole problem, but has contributed part of a solution: "This [working within a single culture] makes it harder to generalise the results, although it also strengthens the contribution of this research since not much is known about stereotypes on computer scientists in the Netherlands" [12].
- *Limited setting? Mitigated.* This dialogue also contains a strong response, providing evidence that the setting, while not universal, is still representative of something larger than itself: "Our results are based on course offerings from a single university, which may limit the generalizability of our findings. However, our program follows the ACM Curriculum guidelines [. . .] and is therefore similar to the curricula used at many institutions." [37].

This example illustrates a situation where the treatment and setting interact, and the interaction has mixed results:

- *Treatment-setting interaction? Trade-off.* "This condition [that students were required to work alone] was necessary for experimental integrity but is not ecologically valid for many classroom lab environments" [55].

In this example, a paper whose data were gathered during the COVID-19 pandemic [4] chose a "Preliminary/suggestive results" response:

- *Treatment-history interaction? Preliminary/suggestive results*: "Although this study was administered during an incredibly difficult time, the work is relevant to supporting what may be a growing number of online courses delivered under so-called normal circumstances in ramping up and supporting service learning experiences" [4].

5.3.2 Participants. Here, we present challenge-response dialogues related to five of the challenges involving studies with human participants.

In one example, the selection of participants may have interacted with the treatment:

- *Treatment-selection interaction? Don't know.* "A second limitation of this study is that while the TEC Rubric was designed for a range of educational decision makers, this study only included teachers; thus, we do not know whether similar patterns of use would be observed in other educational decision makers (e.g., administrators)" [11].

One paper combines three different responses to "Volunteer bias":

- *Volunteer bias? Don't know, Preliminary/suggestive results, and Future work.* "That being the case [that the students were self-selected], we cannot say with certainty whether the results produced with this subset of students would match what would be observed if the entirety of both CS1 classes had taken part in the study [...] [W]e consider this to be only a first study [...] Future research is needed" [46].

In response to "Small participation/response rate", most of the quantitative papers looked to "Future work". One qualitative paper expressed a different perspective, however:

- *Small participation/response rate? Not a problem.* They had a "small sample of participating teachers" but had done a purposive sample and "specifically selected teachers with experience in teaching algorithms" [57].

This example addresses "Demand characteristics" (that students would give answers to please the interviewer):

- *Demand characteristics? Mitigated.* "Specifically, students might alter their responses in order to please the instructor and the TA [...] To mitigate such threats, we enforced anonymity in both of our surveys. [...] [T]he primary class instructor was not present during the surveys. The students were also assured that their answers [...] would not impact their grades" [78].

In response to a challenge of "Self-reported data", one paper offered a "Mitigated" response, explaining that direct observations had been collected as well [22]. Another paper identified the trade-off between measuring experience by self-report instead of a programming test, trading objectivity for time:

- *Self-reported data? Trade-off.* "Of course, a programming test would have been a more objective way of measuring experience, but that would have taken more time for the participants [10].

5.3.3 Researchers. In response to a challenge of "Rater bias", one paper offered this "Mitigated" response:

- *Rater bias? Mitigated.* "Another threat to our study's internal validity might result from the fact that one of the main co-authors of the paper is the main instructor of the class. [...] However, as mentioned earlier, all assignments were graded by the instructor and the TA according to a predefined rubric [...]" [78].

5.3.4 Instruments and materials. In response to a challenge to "Instrument Design/Presentation", one paper offered the mitigation that they had tested the materials in advance:

- *Instrument Design/Presentation? Mitigated.* "Even though we tested the pictures in the pilot, it should be noted that the pictures we choose might result in different associations than the one we were aiming at" [12].

5.3.5 Implementing the study design. In response to a challenge of "Unreliable Implementation", here is an example of a mitigation for having different lecturers teaching different groups in a study:

- *Unreliable Implementation? Mitigated.* "Regarding the reliability of the treatment implementation, both lecturers were informed about the study design, their responsibilities and the expected behavior from them" [54].

5.3.6 Data. Because they are under the researchers' control, "Small sample" challenges are difficult to answer. Some papers leave them unanswered, and some suggest future work. In some cases, responses go further, however, citing support from the literature. We found one example of both:

- *Small sample? No problem, plus Future work.* " [T]his study only included four or five observations for each of the participant's computing and collaborative behaviors. Although this number of observations was consistent with recommendations for C-COI and most CSCL studies examine what make single collaborative experiences successful or unsuccessful [...] additional observations [...] could lead to a fuller understanding [...]" [32].

Here is an example of an argument that "Incomplete data" is not a problem:

- *Incomplete data? Mitigated.* "We also would like to emphasize not all students submitted a solution for every homework assignment [...] Due to the overall high number of submissions for each assignment, we had sufficient data to conduct an analysis that could create valuable insights" [21].

Another illustration of an argument that a challenge is not a problem, in response to a (possibly) "Nonrepresentative sample" (Danish high school students in a gender study):

- *Nonrepresentative sample? Not a problem.* "About three-fifths were women and two-fifths were men which coincides with the gender composition of Danish high schools" [10].

5.3.7 *Validity of statistical conclusions.*

This paper gives an example of a response to an "Omitted variable bias" challenge:

- *Omitted variable bias? Future work.* "Potentially relevant constructs are missing from this study, including mathematics anxiety and working memory [. . .] Although this initial set of studies could not encompass all potentially relevant factors, future research on how such factors are related to bold problem solving is warranted" [48].

5.3.8 *Construct-related evidence for validity.*

An example response to "Inadequate measure of target construct" (Analysis based on written evaluations):

- *Inadequate measure of target construct? Trade-off.* "[W]e do not have insight into the specific motivations of the teachers or how they felt beyond what they wrote. While we think this a limitation, we also think this is an authentic task as it is common for such evaluations to result in a written document detailing the results of the evaluation [11].

5.4 Relation of challenges and research phases

As mentioned in Section 4.2, we coded each challenge with the study phase from which it—according to our interpretation of the authors' account—arose; in total, we found 333 challenges. While not a central part of answering our research questions, the distribution of challenges and phases is of independent interest, and we report the most frequent co-occurrences.

"Experimental design" Of the 216 challenges that arose from the design phase, "Instrument design/presentation" accounted for 39, "Limited setting" for 26, and "Inadequate measure of target construct" for 20 challenges. Almost a quarter of these 216 challenges, 50 challenges, had no response at all. There was a wide distribution of these 50 challenges, with "Instrument design/presentation" (9), "Treatment-selection interaction" (9), and "Limited setting" (8) being the three most frequent ones.

"Experimental execution" A total of 41 challenges was coded as arising from the execution phase. The most frequent of these were "Unreliable implementation" (17), "Treatment-history interaction" (6), and "Small participation/response rate". Of these 41 challenges 16 were not responded to, with "Unreliable implementation" (9), "Treatment-history interaction" (4), and "Instrument design/presentation" (2) being the most frequent ones.

"Data analysis" We coded 31 challenges as arising from the analysis phase. Of these, five related to "Researcher bias" and three challenges each were coded as "Small sample", "Small participation/response rate", "Incomplete data", and "Inadequate measure of target construct". In total, eight challenges arising from this phase were not responded to, but only two types ("Omitted variable bias" and "Small sample") occurred more than once.

"Interpretation" The 43 challenges we coded as arising from the interpretation phase were spread out across more than 20 different types of challenges, only "Omitted variable bias" and "Inadequate measure of target construct" occurred more than twice. Six challenges attributed to this phase were not responded to, but each of them was unique.

The above summary indicates that most of the challenges seem to arise from the "Experimental design" phase of a research study. This is not at all surprising for two reasons: First, as mentioned in Section 2, research designs geared towards mitigating or even eliminating certain challenges are bound to be vulnerable to others [76, Ch. 2]. Second, while even well-funded clinical studies suffer from a tension between soundness and ecological validity [30], running extensive pilot studies which might be suited to eliminate certain challenges arising from a research design is very uncommon in classroom-sized studies which account for the vast majority of the research in our dataset.

6 THREATS AND LIMITATIONS TO THIS PAPER

Nonrepresentative sample? Not a problem: it was a purposive sample. We identified venues that would best illustrate methodological points such as threats to validity, included a venue from mathematics education for comparison, and selected recent data, from the twelve months that ended immediately before we began the analysis (2020–21). Other journals or conferences might yield different results. The three computing-education venues are among the most research-oriented venues in computing education, however. In addition, as seen in Section 5, they did raise a wide variety of challenges and potential responses to those challenges.

Researcher bias? Yes, but mitigated. The researchers all have expertise in reading and analyzing CER papers, but other researchers, influenced by their own experiences, might interpret these papers differently. To minimize this influence, we reflected on our potential biases and took measures to mitigate them, for example working with anonymized excerpts from the 77 papers during most of the analysis phase. We followed a careful coding process and provided a detailed description of that process in Section 4. Finally, we have identified our dataset precisely, enabling others researchers to replicate the work.

Incomplete data? Possibly, but mitigated. We may have failed to identify discussions of threats or limitations within our dataset, if they were not included in a named section or subsection. As noted in Section 4, however, we read each paper and also searched on terms that were likely to occur if any discussion of threats or limitations was present. Of the 23 papers that did not contain a relevantly named section, we did identify 13 that had discussions elsewhere in the paper. Again, the identification of the dataset enables other researchers to replicate the work.

A comment on the impact of the COVID-19 pandemic. One might argue that 2020–2021 might not have been a representative time. We did notice a number of references in ICER papers to data collected during the COVID-19 pandemic ("Treatment-history interaction"). In contrast, due to review cycles and backlogs, journal papers published in 2021 are very likely to report on research in a pre-pandemic setting. Other than "Treatment-history interaction", the challenges mentioned in ICER matched those mentioned in the journals. We have identified the papers in our dataset and endeavored to include sufficient examples here to enable readers to decide for themselves whether this is the case.

When considering whether to look at papers from a pre-COVID year instead, we weighed these concerns against the effects of an important change in the ICER reviewing process: Starting in 2021, ICER introduced "conditional accept" decisions, effectively a very brief revision cycle. Since this revision cycle provides an opportunity for, among other things, responding to threats and limitations raised by reviewers, we decided that the papers from 2021 would provide deeper insights than those from earlier years.

7 DISCUSSION

The fact that talking about threats and limitations has become common but still is not ubiquitous is a concern that is shared with other communities; see, e.g., [7, 17, 29, 31, 75]. Other disciplines raise concerns about possible unreported threats, based on the relatively small number of threats reported in the average paper; see, e.g., [75]. Unreported threats can be difficult to identify, but Lenarduzzi et al. [40] provide an example of how it can be done in specific situations.

We talk about threats and limitations in a way that is not so different from mathematics education, a discipline with a longer tradition. All the challenges we found in JRME were also identified in one or more of the CER venues. The notable difference between JRME and the CER venues was that JRME had no examples of the "Unreliable implementation" challenge. Examining these challenges in context revealed that almost all of these challenges occurred for studies in large undergraduate classes with multiple sections. Since most of the JRME papers were smaller-scale studies, this might explain the absence of "Unreliable implementation" challenges.

Comparing journals and conferences, our dataset included three journals (CSEJ, TOCE, and JRME) and one conference, ICER. The reviewing standards for the three CER venues are quite similar, and ICER, like the journals (and unlike mathematics education conferences), publishes proceedings containing the text of the research papers it accepts. ICER now even offers a (very brief) revise-and-resubmit cycle.

There were two notable difference between the journal papers and those in ICER. First, the "Treatment-history interaction" challenge was raised only in relation to the COVID-19 pandemic, which was only discussed in ICER. Given the different durations between submission and publication for journals and conference, it seems likely that the portions of the studies presented in the journals that would have been affected, such as data gathering, were completed before the pandemic.

Second, ICER's ratio of "Not a problem" and "We don't know" responses to papers was twice that of any of the three journals. This may reflect a difference between conferences and journals; even though ICER now offers a journal-like revision cycle, it is very brief (typically about a week). In general, there has been a trend for computing education conferences to become more like journals in recent years, but examining additional conferences in the field, such as SIGCSE, ITiCSE, Koli, or LaTiCE, might help to characterize the difference between conferences and journals.

Because our study is the first of its kind in computing education research, its scope is designed to be wide. Like Feldt and Magazinius [17] and Brutus et al. [7], but unlike the other studies we reviewed, we also considered qualitative work, reflecting the breadth of different study designs and research methodologies in computing education. Other disciplines with longer traditions include both broad surveys and deeper, more focused investigations of specific threats or threats to particular types of study. (See, e.g., [3, 40, 62, 79]).

Our methodology was also different. The broadly scoped studies that we reviewed used a small, established top-level set of codes, such as "internal validity", "external validity", "construct validity", and "statistical conclusion validity". In contrast, we used a combination of deductive and inductive analysis to develop a hierarchy of codes, as described in Section 4. Only on the lowest level of abstraction of this hierarchy can the connection between challenges and responses be made in a concrete enough way to be instructive. Doing so addresses the concern voiced by [17] that authors might refrain from using established (high-level) terminology because "these terms are not more directly linked to the actual elements of the studies" [17, p. 378].

Most importantly, however, other methodological reviews, focused on evaluating the quality of threats discussions, look only at responses that mitigate or prevent a threat. Our focus, by contrast, is on examining *how* the community talks about challenges, rather than evaluating how well it is done. Because of this different focus, we present a range of responses in addition to mitigation and prevention, and, by looking at all of these responses, we observe a more general pattern. Some papers leave some (or all) of their challenges unanswered, but in general, we find patterns of challenge and response, in which authors engage in a dialogue about the threats to and limitations of their work.

8 CONCLUSIONS AND FUTURE WORK

Despite the value of a good threats and limitations section, it can be difficult to write: it requires the authors to openly talk about the weaknesses in a paper. While this reflects good academic practice and is indicative of a mastery of methodology, some authors seem to be reluctant to do so [63], leaving the reader to identify the threats, if any, from a clear description of the work done. With our study, we have set out to understand the specifics of challenges and responses in computing education research papers.

Considering our research questions: for RQ1 (What challenges?) the challenges raised in the CER literature include many of the "standard" challenges presented in Taylor [76] and the APA Dictionary [2] that we identified deductively. They also include a number of inductively-identified challenges that do not appear in those sources–notably "Incomplete data" and "Instrument design/presentation"—that are common in our CER papers, but also appear in the JRME data.

For RQ2 (What responses?) the responses here were identified inductively by reading and interpreting the texts. Most of the challenges in the papers have one or more responses, and generally explain why the findings are valid notwithstanding the challenges. It is notable that there are a fairly large number of challenges without responses (74/279) in the CER papers, but very few in JRME (3/54), while JRME has more "Future work" responses.

Overall, we found that researchers approach the threats section in two ways: as information to guide the reader in reading, extending, or applying the study presented; and as challenges from a hypothetical audience. Either way, what we see is a pattern of challenge and response. In the vast majority of cases, authors engage with challenges. Most of the time, they avoid simply listing a set of challenges with no response and instead strike a balance, describing challenges but also answering them, at least attempting to do so.

In summary, the contributions of this paper include:

- providing a survey of the discussion of threats and limitations in computing education research;
- linking the threats found to the more abstract terminology used in the theoretical literature;
- contrasting the threats and responses with those found in a math education venue, and the secondary study of threats to validity in computing education with similar secondary studies in other fields;
- giving concrete examples showing the range of threats discussed in our discipline and possible responses to those threats; and
- illustrating the way in which threats and limitations are discussed, in a pattern of challenge and response.

In addition, we hope to have made it easier for authors to write threats sections, first, by providing examples of the range of threats and responses found in computing-education papers, and second, by framing the discussion of threats—which is inevitable, as some challenges occur in every study—as a dialogue between challenge and response.

This paper opens up various possibilities for future work. Our study could be replicated in other venues and years. Studies of challenges to particular aspects of computing education could be done, for example challenges to studies in large introductory classes, to qualitative studies, or to studies involving artifacts. Also, the small but observable differences between the CER papers and the papers from JRME focused on primary education and teachers suggest looking into whether challenges and their response differ by the educational level of the study environment as computing in primary education becomes more and more prevalent.

Finally, we hope that the challenge-response dialogues identified here might serve as a starting point for a deeper conversation about threats and limitations. Which are fatal? Which can be mitigated and how? Which are sometimes not a problem—and what arguments that a threat is not a problem are convincing? Which flaws warrant serious revisions before submission (by the authors) or rejection (by reviewers), and which can be fixed? This paper does not prescribe answers to these questions, but we hope to start a discussion that will move the community closer to some answers.

ACKNOWLEDGMENTS

We are indebted to the reviewers, and also to Tony Clear, Anna Eckerdal, Josh Tenenberg, Lynda Thomas, and Carol Zander, for providing most helpful pointers to literature outside computing education research. Also, we thank Phil Steinhorst for providing the initial version of the program code to visualize the segments and their codes.

REFERENCES

[1] *Efthimia Aivaloglou and Anna van der Meulen. 2021. An Empirical Study of Students' Perceptions on the Setup and Grading of Group Programming Assignments. *ACM Transactions on Computing Education* 21, 3 (Sept. 2021), 17.1–17.22. https://doi.org/10.1145/3440994

[2] American Psychological Association. 2023. APA Dictionary of Psychology. https://dictionary.apa.org. Accessed: 2023-03-11.

[3] Apostolos Ampatzoglou, Stamatia Bibi, Paris Avgeriou, Marijn Verbeek, and Alexander Chatzigeorgiou. 2019. Identifying, categorizing and mitigating threats to validity in software engineering secondary studies. *Information and Software Technology* 106 (2019), 201–230. https://doi.org/10.1016/j.infsof.2018.10.006

[4] *Lecia J. Barker, Amy Voida, and Vaughan Nagy. 2021. Service Interruption: Managing Commitment to Community Partners During a Crisis. In *ICER 2021: Proceedings of the 17th ACM Conference on International Computing Education Research*, Amy J. Ko, Jan Vahrenhold, Renée McCauley, and Matthias Hauswirth (Eds.). ACM Press, New York, NY, 81–91. https://doi.org/10.1145/3446871.3469756

[5] *Jennifer M. Blaney. 2020. Gender and leadership development in undergraduate computing: a closer look at women's leadership conceptions. *Computer Science Education* 30, 4 (2020), 469–499. https://doi.org/10.1080/08993408.2020.1816769

[6] *Nicholas A. Bowman, Lindsay Jarratt, KC Culver, and Alberto M. Segre. 2021. The Impact of Pair Programming on College Students' Interest, Perceptions, and Achievement in Computer Science. *ACM Transactions on Computing Education* 21, 3 (Sept. 2021), 19.1–19.19. https://doi.org/10.1145/3440759

[7] Stéphane Brutus, Herman Aguinis, and Ulrich Wassmer. 2013. Self-Reported Limitations and Future Directions in Scholarly Reports: Analysis and Recommendations. *Journal of Management* 39, 1 (January 2013), 48–75. https://doi.org/10.1177/0149206312455245

[8] Donald T. Campbell and Julian C. Stanley. 1966. *Experimental and Quasi-Experimental Designs for Research.* Houghton Mifflin, Boston, MA.

[9] *Chen Chen, Jane M. Kang, Gerhard Sonnert, and Philip M. Sadler. 2021. High School Calculus and Computer Science Course Taking as Predictors of Success in Introductory College Computer Science. *ACM Transactions on Computing Education* 21, 1 (March 2021), 6.1–6.21. https://doi.org/10.1145/3433169

[10] *Ingrid Maria Christensen, Melissa Høegh Marcher, Paweł Grabarczyk, Therese Graversen, and Claus Brabrand. 2021. Computing Educational Activities Involving People Rather Than Things Appeal More to Women (Recruitment Perspective). In *ICER 2021: Proceedings of the 17th ACM Conference on International Computing Education Research*, Amy J. Ko, Jan Vahrenhold, Renée McCauley, and Matthias Hauswirth (Eds.). ACM Press, New York, NY, 127–144. https://doi.org/10.1145/3446871.3469758

[11] *Merijke Coenraad, Connor Hopcraft, Jane Jozefowicz, Diana Franklin, Jen Palmer, and David Weintrop. 2021. Helping teachers make equitable decisions: effects of the TEC Rubric on teachers' evaluations of a computing curriculum. *Computer Science Education* 31, 3 (2021), 400–429. https://doi.org/10.1080/08993408.2020.1788862

[12] *Shirley de Wit, Felienne Hermans, and Efthimia Aivaloglou. 2021. Children's Implicit and Explicit Stereotypes on the Gender, Social Skills, and Interests of a Computer Scientist. In *ICER 2021: Proceedings of the 17th ACM Conference on International Computing Education Research*, Amy J. Ko, Jan Vahrenhold, Renée McCauley, and Matthias Hauswirth (Eds.). ACM Press, New York, NY, 239–251. https://doi.org/10.1145/3446871.3469753

[13] *Leslie Dietiker and Andrew S. Richman. 2021. How Textbooks Can Promote Inquiry: Using a Narrative Framework to Investigate the Design of Mathematical Content in a Lesson. *Journal for Research in Mathematics Education* 52, 3 (May 2021), 301—-331. https://doi.org/10.5951/jresematheduc-2020-0318

[14] Dennis Drotar. 2009. Editorial: How to Write an Effective Results and Discussion for the Journal of Pediatric Psychology. *Journal of Pediatric Psychology* 34, 4 (May 2009), 339–343. https://doi.org/10.1093/jpepsy/jsp014

[15] *Darrell Earnest and John Chandler. 2021. Making Time: Words, Narratives, and Clocks in Elementary Mathematics. *Journal for Research in Mathematics Education* 52, 4 (July 2021), 407–443. https://doi.org/10.5951/jresematheduc-2021-0020

[16] *Matthew Heinsen Egan and Chris McDonald. 2021. An evaluation of SeeC: a tool designed to assist novice C programmers with program understanding and debugging. *Computer Science Education* 31, 3 (2021), 340–373. https://doi.org/10.1080/08993408.2020.1777034

[17] Robert Feldt and Ana Magazinius. 2010. Validity threats in empirical software engineering research-An initial survey. In *Proceedings of the 22nd International Conference on Software Engineering and Knowledge Engineering (SEKE.* KSI Research Inc., Pittsburgh, PA, 374–379. https://ksiresearchorg.ipage.com/seke/Proceedings/seke/SEKE2010_Proceedings.pdf(Accessed21May2023)

[18] Sally Fincher and Marian Petre. 2004. *Computer Science Education Research.* Taylor and Francis, London.

[19] *James Finnie-Ansley, Paul Denny, and Andrew Luxton-Reilly. 2021. A Semblance of Similarity: Student Categorisation of Simple Algorithmic Problem Statements. In *ICER 2021: Proceedings of the 17th ACM Conference on International Computing Education Research*, Amy J. Ko, Jan Vahrenhold, Renée McCauley, and Matthias Hauswirth (Eds.). ACM Press, New York, NY, 198–212. https://doi.org/10.1145/

3446871.3469745

[20] *Max Fowler, Binglin Chen, and Craig Zilles. 2021. How should we 'Explain in plain English'? Voices from the Community. In *ICER 2021: Proceedings of the 17th ACM Conference on International Computing Education Research*, Amy J. Ko, Jan Vahrenhold, Renée McCauley, and Matthias Hauswirth (Eds.). ACM Press, New York, NY, 69–80. https://doi.org/10.1145/3446871.3469738

[21] *Christiane Frede and Maria Knobelsdorf. 2021. A differentiated picture of student performance in introductory courses to theory of computation. *Computer Science Education* 31, 3 (2021), 315–339. https://doi.org/10.1080/08993408.2020.1809946

[22] *Joanna Goode, Allison Ivey, Stephany RunningHawk Johnson, Jean J. Ryoo, and Christine Ong. 2021. Rac(e)ing to computer science for all: how teachers talk and learn about equity in professional development. *Computer Science Education* 31, 3 (2021), 374–399. https://doi.org/10.1080/08993408.2020.1804772

[23] Sue Greener. 2018. Research limitations: the need for honesty and common sense. *Interactive Learning Environments* 26, 5 (2018), 567–568. https://doi.org/10.1080/10494820.2018.1486785

[24] *Gregor Große-Bölting, Dietrich Gerstenberger, Lara Gildehaus, Andreas Mühling, and Carsten Schulte. 2021. Identity in K-12 Computer Education Research: A Systematic Literature Review. In *ICER 2021: Proceedings of the 17th ACM Conference on International Computing Education Research*, Amy J. Ko, Jan Vahrenhold, Renée McCauley, and Matthias Hauswirth (Eds.). ACM Press, New York, NY, 169–183. https://doi.org/10.1145/3446871.3469757

[25] *Joonas Häkkinen, Petri Ihantola, Matti Luukkainen, Antti Leinonen, and Juho Leinonen. 2021. Persistence of Time Management Behavior of Students and Its Relationship with Performance in Software Projects. In *ICER 2021: Proceedings of the 17th ACM Conference on International Computing Education Research*, Amy J. Ko, Jan Vahrenhold, Renée McCauley, and Matthias Hauswirth (Eds.). ACM Press, New York, NY, 92–100. https://doi.org/10.1145/3446871.3469767

[26] *Sally Hamouda, Stephen H. Edwards, Hicham G. Elmongui, Jeremy V. Ernst, and Clifford A. Shaffer. 2020. BTRecurTutor: a tutorial for practicing recursion in binary trees. *Computer Science Education* 30, 2 (2020), 216–248. https://doi.org/10.1080/08993408.2020.1714533

[27] *Rachelle Haroldson and Dave Ballard. 2021. Alignment and representation in computer science: an analysis of picture books and graphic novels for K-8 students. *Computer Science Education* 31, 1 (2021), 4–29. https://doi.org/10.1080/08993408.2020.1779520

[28] *Mohammed Hassan and Craig Zilles. 2021. Exploring 'reverse-tracing' Questions as a Means of Assessing the Tracing Skill on Computer-based CS 1 Exams. In *ICER 2021: Proceedings of the 17th ACM Conference on International Computing Education Research*, Amy J. Ko, Jan Vahrenhold, Renée McCauley, and Matthias Hauswirth (Eds.). ACM Press, New York, NY, 115–126. https://doi.org/10.1145/3446871.3469765

[29] Esther Helmich, Benjamin C. M. Boerebach, Onyebuchi A. Arah, and Lorelei Lingard. 2015. Beyond limitations: Improving how we handle uncertainty in health professions education research. *Medical Teacher* 37, 11 (2015), 1043–1050. https://doi.org/10.3109/0142159X.2015.1073239

[30] Stephen B. Hulley, Stephen R. Cummings, Warren S. Browner, and Deborah G. Grady. 2013. *Designing Clinical Research* (4 ed.). Wolters Kluwer Health, Philadelphia, PA.

[31] John P.A. Ioannidis. 2007. Limitations are not properly acknowledged in the scientific literature. *Journal of Clinical Epidemiology* 60, 4 (April 2007), 324–329. https://doi.org/10.1016/j.jclinepi.2006.09.011

[32] *Maya Israel, Moon Y. Chung, Quentin M. Wherfel, and Saddeddine Shehab. 2020. A descriptive analysis of academic engagement and collaboration of students with autism during elementary computer science. *Computer Science Education* 30, 4 (2020), 444–468. https://doi.org/10.1080/08993408.2020.1779521

[33] Michael T. Kane. 1992. An Argument-Based Approach to Validity. *Psychological Bulletin* 112, 3 (Nov. 1992), 527–535. https://doi.org/10.1037/0033-2909.112.3.527

[34] Michael T. Kane. 2009. Validating the Interpretation and Uses of Test Scores. In *The Concept of Validity: Revision, New Directions, and Applications*, Robert W. Lissitz (Ed.). Information Age Publishing, Charlotte, NC, Chapter 3, 39–64.

[35] Amy J. Ko and Sally A. Fincher. 2019. A Study Design Process. In *The Cambridge Handbook of Computing Education Research*, Sally A. Fincher and Anthony V. Robins (Eds.). Cambridge University Press, Cambridge, Chapter 4, 81–101. https://doi.org/10.1017/9781108654555.005

[36] *Inah Ko and Patricio Herbst. 2020. Subject Matter Knowledge of Geometry Needed in Tasks of Teaching: Relationship to Prior Geometry Teaching Experience. *Journal for Research in Mathematics Education* 51, 5 (Nov. 2020), 600–630. https://doi.org/10.5951/jresematheduc-2020-0163

[37] *Sophia Krause-Levy, William G. Griswold, Leo Porter, and Christine Alvarado. 2021. The Relationship Between Sense of Belonging and Student Outcomes in CS1 and Beyond. In *ICER 2021: Proceedings of the 17th ACM Conference on International Computing Education Research*, Amy J. Ko, Jan Vahrenhold, Renée McCauley, and Matthias Hauswirth (Eds.). ACM Press, New York, NY, 29–41. https://doi.org/10.1145/3446871.3469748

[38] *Shriram Krishnamurthi and Kathi Fisler. 2021. Developing Behavioral Concepts of Higher-Order Functions. In *ICER 2021: Proceedings of the 17th ACM Conference on International Computing Education Research*, Amy J. Ko, Jan Vahrenhold, Renée McCauley, and Matthias Hauswirth (Eds.). ACM Press, New York, NY, 306–318. https://doi.org/10.1145/3446871.3469739

[39] H. Chad Lane and Kurt VanLehn. 2003. Coached Program Planning: Dialogue-Based Support for Novice Program Design. In *Proceedings of the 34th SIGCSE Technical Symposium on Computer Science Education* (Reno, Nevada, USA) *(SIGCSE '03)*. ACM Press, New York, NY, 148–152. https://doi.org/10.1145/611892.611955

[40] Valentina Lenarduzzi, Oscar Dieste, Davide Fucci, and Sira Vegas. 2021. Towards a Methodology for Participant Selection in Software Engineering Experiments: A Vision of the Future. In *Proceedings of the 15th ACM / IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM)* (Bari, Italy) *(ESEM '21)*. Association for Computing Machinery, New York, NY, USA, Article 35, 6 pages. https://doi.org/10.1145/3475716.3484273

[41] *Lara Letaw, Rosalinda Garcia, Heather Garcia, Christopher Perdriau, and Margaret Burnett. 2021. Changing the Online Climate via the Online Students: Effects of Three Curricular Interventions on Online CS Students' Inclusivity. In *ICER 2021: Proceedings of the 17th ACM Conference on International Computing Education Research*, Amy J. Ko, Jan Vahrenhold, Renée McCauley, and Matthias Hauswirth (Eds.). ACM Press, New York, NY, 42–59. https://doi.org/10.1145/3446871.3469742

[42] Hareton K. N. Leung. 2003. Evaluating the Effectiveness of e-Learning. *Computer Science Education* 13, 2 (2003), 123–136. https://doi.org/10.1076/csed.13.2.123.14201

[43] *Luis A. Leyva. 2021. Black Women's Counter-Stories of Resilience and Within-Group Tensions in the White, Patriarchal Space of Mathematics Education. *Journal for Research in Mathematics Education* 52, 2 (March 2021), 115–151. https://doi.org/10.5951/jresematheduc-2020-0027

[44] Yvonna S. Lincoln and Egon G. Guba. 1985. *Naturalistic Inquiry.* SAGE Publications, Newbury Park, CA.

[45] Lorelei Lingard. 2015. The art of limitations. *Perspectives on Medical Education* 4, 3 (May 2015), 136–137. https://doi.org/10.1007/S40037-015-0181-0

[46] *Alex Lishinski and Joshua Rosenberg. 2021. All the Pieces Matter: The Relationship of Momentary Self-efficacy and Affective Experiences with CS1 Achievement and Interest in Computing. In *ICER 2021: Proceedings of the 17th ACM Conference on International Computing Education Research*, Amy J. Ko, Jan Vahrenhold, Renée McCauley, and Matthias Hauswirth (Eds.). ACM Press, New York, NY, 252–265. https://doi.org/10.1145/3446871.3469740

[47] *Alex Lishinski and Aman Yadav. 2021. Self-evaluation Interventions: Impact on Self-efficacy and Performance in Introductory Programming. *ACM Transactions on Computing Education* 21, 3 (Sept. 2021), 23.1–23.28. https://doi.org/10.1145/3447378

[48] *Sarah Theule Lubienski, Colleen M. Ganley, Martha B. Makowski, Emily K. Miller, and Jennifer D. Timmer. 2021. "Bold Problem Solving": A New Construct for Understanding Gender Differences in Mathematics. *Journal for Research in Mathematics Education* 52, 1 (Jan. 2021), 12–61. https://doi.org/10.5951/jresematheduc-2020-0136

[49] Joseph A. Maxwell. 1992. Understanding and Validity in Qualitative Research. *Harvard Educational Review* 62, 3 (Fall 1992), 279–301. https://doi.org/10.17763/haer.62.3.8323320856251826

[50] Nora McDonald, Sarita Schoenebeck, and Andrea Forte. 2019. Reliability and Inter-rater Reliability in Qualitative Research: Norms and Guidelines for CSCW and HCI Practice. *Proceedings of the ACM on Human-Computer Interaction* 3 (Nov. 2019), 72:1–72:23. Issue CSCE. https://doi.org/10.1145/3359174

[51] Monica M. McGill, Adrienne Decker, and Zachary Abbott. 2018. Improving Research and Experience Reports of Pre-College Computing Activities: A Gap Analysis. In *Proceedings of the 49th ACM Technical Symposium on Computer Science Education*, Tiffany Barnes, Daniel D. Garcia, Elizabeth K. Hawthorne, and Manuel A. Pérez-Quiñones (Eds.). ACM Press, New York, NY, 964–969. https://doi.org/10.1145/3159450.3159481

[52] *Daphne Miedema, Efthimia Aivaloglou, and George Fletcher. 2021. Identifying SQL Misconceptions of Novices: Findings from a Think-Aloud Study. In *ICER 2021: Proceedings of the 17th ACM Conference on International Computing Education Research*, Amy J. Ko, Jan Vahrenhold, Renée McCauley, and Matthias Hauswirth (Eds.). ACM Press, New York, NY, 355–367. https://doi.org/10.1145/3446871.3469759

[53] *Alexandra Milliken, Veronica Cateté, Ally Limke, Isabella Gransbury, Hannah Chipman, Yihuan Dong, and Tiffany Barnes. 2021. Exploring and Influencing Teacher Grading for Block-based Programs through Rubrics and the GradeSnap Tool. In *ICER 2021: Proceedings of the 17th ACM Conference on International Computing Education Research*, Amy J. Ko, Jan Vahrenhold, Renée McCauley, and Matthias Hauswirth (Eds.). ACM Press, New York, NY, 101–114. https://doi.org/10.1145/3446871.3469762

[54] *Miguel Ehécatl Morales-Trujillo and Gabriel Alberto García-Mireles. 2021. Gamification and SQL: An Empirical Study on Student Performance in a Database Course. *ACM Transactions on Computing Education* 21, 1 (March 2021), 3.1–3.29. https://doi.org/10.1145/3427597

[55] *Briana B. Morrison, Lauren E. Margulieux, and Adrienne Decker. 2020. The curious case of loops. *Computer Science Education* 30, 2 (2020), 127–154. https://doi.org/10.1080/08993408.2019.1707544

[56] Amadeu Anderlin Neto and Tayana Conte. 2013. A Conceptual Model to Address Threats to Validity in Controlled Experiments. In *Proceedings of the 17th International Conference on Evaluation and Assessment in Software Engineering* (Porto de Galinhas, Brazil) *(EASE '13)*. Association for Computing Machinery, New York, NY, USA, 82–85. https://doi.org/10.1145/2460999.2461011

[57] *Jacqueline Nijenhuis-Voogt, Durdane Bayram-Jacobs, Paulien C. Meijer, and Erik Barendsen. 2021. Omnipresent yet elusive: Teachers' views on contexts for teaching algorithms in secondary education. *Computer Science Education* 31, 1 (2021), 30–59. https://doi.org/10.1080/08993408.2020.1783149

[58] Alannah Oleson, Benjamin Xie, Jean Salac, Jayne Everson, J. Megumi Kivuva, and Amy J. Ko. 2022. A Decade of Demographics in Computing Education Research: A Critical Review of Trends in Collection, Reporting, and Use. In *ICER '22: Proceedings of the 2022 ACM Conference on International Computing Education Research – Volume 1*, Jan Vahrenhold, Kathi Fisler, Matthias Hauswirth, and Diana Franklin (Eds.). ACM Press, New York, NY, 323–343. https://doi.org/10.1145/3501385.3543967

[59] *Miranda C. Parker, Mark Guzdial, and Allison Elliott Tew. 2021. Uses, Revisions, and the Future of Validated Assessments in Computing Education: A Case Study of the FCS1 and SCS1. In *ICER 2021: Proceedings of the 17th ACM Conference on International Computing Education Research*, Amy J. Ko, Jan Vahrenhold, Renée McCauley, and Matthias Hauswirth (Eds.). ACM Press, New York, NY, 60–68. https://doi.org/10.1145/3446871.3469744

[60] *Marios Pittalis, Demetra Pitta-Pantazi, and Constantinos Christou. 2020. Young Students' Functional Thinking Modes: The Relation Between Recursive Patterning, Covariational Thinking, and Correspondence Relations. *Journal for Research in Mathematics Education* 51, 5 (Nov. 2020), 631–674. https://doi.org/10.5951/jresematheduc-2020-0164

[61] *Seth Poulsen, Mahesh Viswanathan, Geoffrey L. Herman, and Matthew West. 2021. Evaluating Proof Blocks Problems as Exam Questions. In *ICER 2021: Proceedings of the 17th ACM Conference on International Computing Education Research*, Amy J. Ko, Jan Vahrenhold, Renée McCauley, and Matthias Hauswirth (Eds.). ACM Press, New York, NY, 157–168. https://doi.org/10.1145/3446871.3469741

[62] James Price, Judy Murnan, Joseph A. Drake, Jamie Dimmig, and Mary Hayes. 2004. Mail Survey Return Rates Published in Health Education Journals: An Issue of External Validity. *American Journal of Health Education* 35 (2004), 19–23. Issue 1. https://doi.org/10.1080/19325037.2004.10603600

[63] James H. Price and Judy Murnan. 2004. Research Limitations and the Necessity of Reporting Them. *American Journal of Health Education* 35, 2 (2004), 66–67. https://doi.org/10.1080/19325037.2004.10603611

[64] Milo A. Puhan, Elie A. Akl, Dianne Bryant, Feng Xie, Giovanni Apolone, and Gerben ter Riet. 2012. Discussing study limitations in reports of biomedical studies – the need for more transparency. *Health and Quality of Life Outcomes* 10, Article 23 (2012), 4 pages. https://doi.org/10.1186/1477-7525-10-23

[65] Justus J. Randolph. 2007. *Computer Science Education Research at the Crossroads: A Methodological Review of Computer Science Education Research: 2000–2005.* Ph. D. Dissertation. Department of Psychology, Utah State University, Logan, UT.

[66] Justus J. Randolph, George Julnes, Erkki Sutinen, and Steve Lehman. 2008. A methodological review of computer science education research. *Journal of Information Technology Education* 7 (2008), 135–162. https://doi.org/10.28945/183

[67] Paula T. Ross and Nikki L. Bibler Zaidi. 2019. Limited by our limitations. *Perspectives on Medical Education* 8, 4 (July 2019), 261–264. https://doi.org/10.1007/S40037-019-00530-X

[68] *Jennifer Ruef. 2021. How Ms. Mayen and Her Students Co-Constructed Good-at-Math. *Journal for Research in Mathematics Education* 52, 2 (March 2021), 152–188. https://doi.org/10.5951/jresematheduc-2020-0264

[69] *Evthokia Stephanie Saclarides and Sarah Theule Lubienski. 2021. Teachers' Mathematics Learning Opportunities During One-on-One Coaching Conversations. *Journal for Research in Mathematics Education* 52, 3 (May 2021), 257–300. https://doi.org/10.5951/jresematheduc-2020-0092

[70] *Jean Salac, Cathy Thomas, Chloe Butler, and Diana Franklin. 2021. Investigating the Role of Cognitive Abilities in Computational Thinking for Young Learners. In *ICER 2021: Proceedings of the 17th ACM Conference on International Computing Education Research*, Amy J. Ko, Jan Vahrenhold, Renée McCauley, and Matthias Hauswirth (Eds.). ACM Press, New York, NY, 2–17. https://doi.org/10.1145/3446871.3469746

[71] *Adrian Salguero, William G. Griswold, Christine Alvarado, and Leo Porter. 2021. Understanding Sources of Student Struggle in Early Computer Science Courses. In *ICER 2021: Proceedings of the 17th ACM Conference on International Computing Education Research*, Amy J. Ko, Jan Vahrenhold, Renée McCauley, and Matthias Hauswirth (Eds.). ACM Press, New York, NY, 319–333. https://doi.org/10.1145/3446871.3469755

[72] Kate Sanders, Judy Sheard, Brett A. Becker, Anna Eckerdal, Sally Hamouda, and Simon. 2019. Inferential Statistics in Computing Education Research: A Methodological Review. In *Proceedings of the 2019 ACM Conference on International Computing Education Research, ICER 2019*, Robert McCartney, Andrew Petersen, Anthony V. Robins, and Adon Moskal (Eds.). ACM Press, New York, NY, 177–185. https://doi.org/10.1145/3291279.3339408

[73] *Sue Sentance and Jane Waite. 2021. Teachers' Perspectives on Talk in the Programming Classroom: Language as a Mediator. In *ICER 2021: Proceedings of the 17th ACM Conference on International Computing Education Research*, Amy J. Ko, Jan Vahrenhold, Renée McCauley, and Matthias Hauswirth (Eds.). ACM Press, New York, NY, 266–280. https://doi.org/10.1145/3446871.3469751

[74] Simon, Brett A. Becker, Sally Hamouda, Robert McCartney, Kate Sanders, and Judy Sheard. 2019. Visual Portrayals of Data and Results at ITiCSE. In *Proceedings of the 2019 ACM Conference on Innovation and Technology in Computer Science Education, ITiCSE 2019*, Bruce Scharlau, Roger McDermott, Arnold Pears, and Mihaela Sabin (Eds.). ACM, New York, NY, 51–57. https://doi.org/10.1145/3304221.3319742

[75] Dag I. K. Sjøberg, Jo Erskine Hannay, Ove Hansen, Vigdis By Kampenes, Amela Karahasanovic, Nils-Kristian Liborg, and Anette C. Rekdal. 2005. A survey of controlled experiments in software engineering. *IEEE Transactions on Software Engineering* 31 (Sept. 2005), 733–753. Issue 9. https://doi.org/10.1109/TSE.2005.97

[76] Catherine S. Taylor. 2012. *Validity and Validation.* Oxford University Press, Oxford.

[77] *Kyle Thayer, Sarah E. Chasins, and Amy J. Ko. 2021. A Theory of Robust API Knowledge. *ACM Transactions on Computing Education* 21, 1 (March 2021), 8.1–8.32. https://doi.org/10.1145/3444945

[78] *Miroslav Tushev, Grant Williams, and Anas Mahmoud. 2020. Using GitHub in large software engineering classes. An exploratory case study. *Computer Science Education* 30, 2 (2020), 155–186. https://doi.org/10.1080/08993408.2019.1696168

[79] Michael T. M. Wang, Mark J. Bolland, and Andrew Grey. 2015. Reporting of limitations of observational research. *JAMA Internal Medicine* 175, 9 (Sept. 2015), 1571–1572. https://doi.org/10.1001/jamainternmed.2015.2147

[80] *Dana Linnell Wanzer, Tom McKlin, Jason Freeman, Brian Magerko, and Taneisha Lee. 2020. Promoting intentions to persist in computing: an examination of six years of the EarSketch program. *Computer Science Education* 30, 4 (2020), 394–419. https://doi.org/10.1080/08993408.2020.1714313

[81] *Zhen Xu, Albert D. Ritzhaupt, Karthikeyan Umapathy, Yang Ning, and Chin-Chung Tsai. 2021. Exploring college students' conceptions of learning computer science: a draw-a-picture technique study. *Computer Science Education* 31, 1 (2021), 60–82. https://doi.org/10.1080/08993408.2020.1783155

[82] *Iman YeckehZaare, Elijah Fox, Gail Grot, Sean Chen, Claire Walkosak, Kevin Kwon, Annelise Hofmann, Jessica Steir, Olivia McGeough, and Nealie Silverstein. 2021. Incentivized Spacing and Gender in Computer Science Education. In *ICER 2021: Proceedings of the 17th ACM Conference on International Computing Education Research*, Amy J. Ko, Jan Vahrenhold, Renée McCauley, and Matthias Hauswirth (Eds.). ACM Press, New York, NY, 18–28. https://doi.org/10.1145/3446871.3469760

[83] *Nick Young and Shriram Krishnamurthi. 2021. Early Post-Secondary Student Performance of Adversarial Thinking. In *ICER 2021: Proceedings of the 17th ACM Conference on International Computing Education Research*, Amy J. Ko, Jan Vahrenhold, Renée McCauley, and Matthias Hauswirth (Eds.). ACM Press, New York, NY, 213–224. https://doi.org/10.1145/3446871.3469743

[84] *Albina Zavgorodniaia, Artturi Tilanterä, Ari Korhonen, Otto Seppälä, Arto Hellas, and Juha Sorva. 2021. Algorithm Visualization and the Elusive Modality Effect. In *ICER 2021: Proceedings of the 17th ACM Conference on International Computing Education Research*, Amy J. Ko, Jan Vahrenhold, Renée McCauley, and Matthias Hauswirth (Eds.). ACM Press, New York, NY, 368–378. https://doi.org/10.1145/3446871.3469747

[85] Gerlese S. Åkerlind. 2005. Variation and commonality in phenomenographic research methods. *Higher Education Research & Development* 24, 4 (2005), 321–334. https://doi.org/10.1080/07294360500284672