A Systematic Literature Review of Empiricism and Norms of Reporting in Computing Education Research Literature

SARAH HECKMAN, North Carolina State University
JEFFREY C. CARVER, University of Alabama
MARK SHERRIFF, University of Virginia
AHMED AL-ZUBIDY, University of Alabama

Context. Computing Education Research (CER) is critical to help the computing education community and policy makers support the increasing population of students who need to learn computing skills for future careers. For a community to systematically advance knowledge about a topic, the members must be able to understand published work thoroughly enough to perform replications, conduct meta-analyses, and build theories. There is a need to understand whether published research allows the CER community to systematically advance knowledge and build theories.

Objectives. The goal of this study is *to characterize the reporting of empiricism in Computing Education Research literature by identifying whether publications include content necessary for researchers to perform replications, meta-analyses, and theory building.* We answer three research questions related to this goal: (RQ1) What percentage of papers in CER venues have some form of empirical evaluation? (RQ2) Of the papers that have empirical evaluation, what are the characteristics of the empirical evaluation? (RQ3) Of the papers that have empirical evaluation, do they follow norms (both for inclusion and for labeling of information needed for replication, meta-analysis, and, eventually, theory-building) for reporting empirical work?

Methods. We conducted a systematic literature review of the 2014 and 2015 proceedings or issues of five CER venues: *Technical Symposium on Computer Science Education* (SIGCSE TS), *International Symposium on Computing Education Research* (ICER), *Conference on Innovation and Technology in Computer Science Education* (ITiCSE), *ACM Transactions on Computing Education* (TOCE), and *Computer Science Education* (CSE). We developed and applied the *CER Empiricism Assessment Rubric* to the 427 papers accepted and published at these venues over 2014 and 2015. Two people evaluated each paper using the *Base Rubric* for characterizing the paper. An individual person applied the other rubrics to characterize the norms of reporting, as appropriate for the paper type. Any discrepancies or questions were discussed between multiple reviewers to resolve.

Results. We found that over 80% of papers accepted across all five venues had some form of empirical evaluation. Quantitative evaluation methods were the most frequently reported. Papers most frequently reported results on interventions around pedagogical techniques, curriculum, community, or tools. There was a split in papers that had some type of comparison between an intervention and some other dataset or baseline. Most papers reported related work, following the expectations for doing so in the SIGCSE and CER community. However, many papers were lacking properly reported research objectives, goals, research questions, or

hypotheses; description of participants; study design; data collection; and threats to validity. These results align with prior surveys of the CER literature.

Conclusions. CER authors are contributing empirical results to the literature; however, not all norms for reporting are met. We encourage authors to provide clear, labeled details about their work so readers can use the study methodologies and results for replications and meta-analyses. As our community grows, our reporting of CER should mature to help establish computing education theory to support the next generation of computing learners.

CCS Concepts: • **General and reference** → **Surveys and overviews**; • **Social and professional topics** → *Computer science education;*

Additional Key Words and Phrases: Systematic literature review, empiricism, computing education research

1 INTRODUCTION

From 2009 to 2015, the number of bachelor's degrees in Computer Science increased by 74 percent, while overall growth across all fields of study only rose by 16 percent [50]. Taulbee's 2019 [75] report shows an increase from a recent minimum of under 10,000 bachelor's degrees conferred in 2009 to over 30,000 degrees in 2019 (see Figure 1(B) [75]). It is undeniable that there has been a rapid increase in interest in computer science courses at higher-education institutions. We see similar demand in K–12 with CSforAll[1] initiatives. Educators have struggled with how to cope with new and exacerbated challenges in computing education—ranging from how to scale up to handle new enrollments to which teaching techniques are best during this time of growth. The COVID-19 pandemic has added additional investigation into the challenges of online instruction.

At the same time, attendance at the ACM SIGCSE Technical Symposium has continued to increase year over year, coinciding with a similar increase in the number of submissions [25, 74]. Computing faculty are eagerly trying new teaching methodologies to address the challenges they face with growth in enrollment and with ever-emerging technologies. Other **computing education research (CER)** venues are also seeing increases. The SIGCSE Board[2] added a fourth conference to its yearly schedule to address the growing need for CS education in parts of the world not served by existing conferences [30]. There are many open questions in computing education that the CER community must answer [9, 33].

While this growth in the CER community is exciting, the real benefit comes when community members report innovations in a manner that allows other researchers to build on them and educators to appropriately apply them in their own contexts. For a community to systematically advance knowledge, its members must be able to understand published work thoroughly enough to perform replications, conduct meta-analyses, and build theories [4, 8, 12, 16, 19, 21, 24, 29, 41, 42, 59]. In education literature, this concept is referred to as *the Scholarship of Teaching and Learning* or SoTL. SoTL is broadly defined as "posing [a problem] about an issue of teaching or learning, study of the

---

[1]See https://www.csforall.org/ for more details.
[2]We use SIGCSE to represent the ACM Special Interest Group in Computer Science Education as a community of computing education researchers and practitioners. We use SIGCSE Technical Symposium (SIGCSE TS) to represent the Technical Symposium on Computer Science Education, which is commonly called SIGCSE. This differentiates the community from the conference.

problem through methods appropriate to the disciplinary epistemologies, application of results to practice, communication of results, self-reflection, and peer review" [8].

In 2004, Fincher and Petre posited that most CER publications lack the type of evidence or replications required for meta-analysis or theory building [16]. Clear also observed that many of the standard research practices used to demonstrate rigor in "traditional," non-human, computing topic areas are ignored or even actively avoided when performing CER [11]. These observations occurred just prior to the transition to what Guzdial and du Boulay [22] call the "modern era" of CER in 2005 with the creation of the **International Computing Education Research Conference (ICER)**. Because this work was more than 10 years ago, there is a need to understand how well more recently published research provides the type of information necessary for the computing education community to systematically advance knowledge. Our work complements recent work in understanding the quality of reporting in CER literature [4, 24, 29, 41, 42, 45] and more broadly [7, 12, 19, 60] to support replication.

Therefore, to provide insight into the current state of CER publications, the goal of this study is:

> **To characterize the reporting of empiricism in Computing Education Research literature by identifying whether publications include content necessary for researchers to perform replications, meta-analyses, and theory building.**

To accomplish this goal, we first defined the *CER Empiricism Assessment Rubric* for analyzing CER papers. The rubric identifies the information necessary for replication, meta-analysis, and theory building. We then applied the rubric to 427 papers published in the *Technical Symposium on Computer Science Education (SIGCSE TS)*, *International Symposium on Computing Education Research (ICER)*, *Conference on Innovation and Technology in Computer Science Education (ITiCSE)*, *ACM Transactions on Computing Education (TOCE),* and *Computer Science Education (CSE)* during 2014 and 2015 to categorize published work and identify whether information needed for replication is present and clearly labeled. We chose these years because they were the most recent editions of the conference when we began our research. In addition, these years coincide with 10 years after the first ICER [1, 2], a conference focused on empirical CER and just as the number of empirical research papers published at the SIGCSE TS began to increase. Therefore, the results of this analysis will serve as a baseline against which we can compare the results of a similar analysis in subsequent years.

The contributions of this article are:

- The *CER Empiricism Assessment Rubric* for evaluating the completeness of the empirical content of CER papers;
- An analysis of the empiricism present in CER papers published during 2014 and 2015 in five CER venues;
- A baseline for future analyses; and
- Overall observations that serve as suggestions for how the CER community can advance scientific reporting standards.

2 RELATED WORK

We explore related work in three ways: (1) CER literature reviews and community reflection; (2) guidelines for reporting empirical work, particularly educational work, outside of computing; and (3) an overview of recommendations for replications. These views of the related literature show the current efforts in transforming and increasing the impact of CER results and transfer into practice

that support learners at a variety of levels and needs. We close the section with a discussion on where our work fits into the broader literature.

2.1 CER Literature Reviews and Community Reflection

There have been many reviews of computing education literature, especially over the past 20 years. Holmboe et al. [28] and Clear [11] provide early reflections on CER and they note the biases in the broader research community, including computing, against (computing) education research as a distinct and robust research area [11, 28]. Supporting the growth and recognition of CER and researchers through literature reviews and community reflection continues the progress made by the CER community in the past 15–20 years.

For this section, our focus is on reviews that categorize or evaluate CER papers and reporting quality, including the use of theories and measurements. We do not include systematic literature reviews on specific topics in computing education if they do not also include some discussion on the quality of or challenges in aggregation due to reporting (e.g., reviews of K–12 [20, 70]; introductory programming [46, 53, 73]; teaching assistants [47]; tools, languages, environments in K–12 [44]; meta-cognition and self-regulated learning [55]; and mapping of theories in CER [69]). Table 1 summarizes the reviews that include categorization or evaluation of CER papers from a quality and reporting perspective. For each review, we provide the reference information, area of focus, the years and venues considered, the number of papers reviewed, and a summary of inclusion/exclusion criteria beyond the area of focus. The text below also includes non-systematic literature review papers; however, Table 1 does not list these papers. We organize the discussion in each subsection in chronological order by publication date to highlight the changes in reporting since the early 2000s.

2.1.1 CER Topics in Literature. We start with an overview of how authors categorized topics in CER publications.

The seminal book by Fincher and Petre [16] [2004] identified 10 core areas for CER and provided an initial definition of the practices and methods of the field [16]. Their work laid the foundation for CER in the next 15 years. Pears et al. [52] [2005] created a taxonomy of four key areas in CER that builds on and groups Fincher's and Petre's [16] initial categorization. The categories included studies in teaching and learning; institutions and educational settings; problems and solutions; and CER as a discipline [52]. From these categories, Pears et al. [52] created a core CER literature including influential, seminal, and synthesis work.

Joy et al. [32] [2009] created a taxonomy to categorize the types of CER publications in 21 journals and 21 conference proceedings from either 2004 or 2005. Their findings demonstrate that conference venues tend to have a technical focus to their proceedings, while journals have a pedagogical focus.

Some survey papers identified gaps in the topics covered by CER literature. Kinnunen et al. [33] [2010] reviewed 67 ICER papers between 2005 and 2009. Their categorization considered eight categories based on a three-layered didactic structure, including students, teachers, and goals/context at a classroom, organization, and societal level. Expanding their corpus of reviewed papers by 13 and considering venues beyond ICER (e.g., ACE, SIGCSE TS, PPIG, ITiCSE, ITiCSE WGR, CSE, Comput. Small Coll.), they found papers reported results at the course-level and focused on student characteristics and process. Eight papers included additional focus areas of students' conceptions on and actions to achieve course goals and content. The literature at that time did not provide adequate coverage of categories around content/goals and teachers.

Simon and others (2007–2020) have conducted extensive classification studies of computing education literature for ACE [64, 65], ICER [66], ITiCSE [67], Koli [63], and NACCQ [68]. The

Table 1. Summary of CER Reviews

Reference	Area of Focus	Years	Venues	# Papers	Inclusion Criteria
Valentine [72] 2004	first-year CS courses including CS1 and CS2	1984–2003	SIGCSE TS	444	CS1, CS2, and CS1/CS2 topics
Simon [63] 2007b	computing education paper classification	2001–2006	Koli	102	full papers
Randolph et al. [56] 2008	methodological properties of reported research	2000–2005	SIGCSE Bulletin, CSE, JCSE, Koli, SIGCSE TS, ITiCSE, ICER, ACE	352	stratified random sample from all articles from listed venues between given years
Simon et al. [66] 2008a	computing education paper classification	2005–2007	ICER	43	full papers
Simon et al. [68] 2008b	computing education paper classification	2000–2007	NACCQ	157	full papers accepted to CompEd track
Simon [64] 2009	computing education paper classification	1996–2008	ACE	328	full papers
Joy et al. [32] 2009	categorizing types of CER literature	2004 and/or 2005	Too many to list	3,598	one proceeding/volume from conference/journal
Sheard et al. [61] 2009	teaching and learning of programming	2005–2008	ICER, ITiCSE, SIGCSE TS, ACE, Koli, NACCQ	164	programming-focused paper classified as an experiment, study, or analysis
Kinnunen et al. [33] 2010	didactic-focus-based categorization	2005–2009	ICER	67	focus on instructional process
Malmi et al. [40] 2010	CER research processes	2005–2009	ICER	72	full papers
Malmi et al. [39] 2014	theories, conceptual models, frameworks of CER literature	2005–2011	ERIC/TOCE, CSE, ICER	308	peer-reviewed full papers
Lishinski et al. [36] 2016	use of educational theory empirical results & methodological rigor	2012–2015	CSE, ICER	136	full papers
Ihantola et al. [29] 2015	educational data mining and learning analytics	2005–2015	SIGCSE TS, ICER, ITiCSE, TOCE, CSE, ACE, EDM, JEDM, Koli, PPIG, TLT	76	open-ended programming problems, programming process, automated data collection and analysis, length >3 pages
Al-Zubidy et al. [5] 2016	empiricism in CS education	2014–2015	SIGCSE TS	162	full papers with empirical results
Luxton-Reilly et al. [37] 2018	introductory programming	2003–2017	ACM full-text collection, IEEE Explore, Science Direct (Elsevier), SpringerLink, Scopus	1,666	introductory programming courses student in computing degrees

(Continued)

Table 1. Continued

Reference	Area of Focus	Years	Venues	# Papers	Inclusion Criteria
McGill et al. [45] 2018	status of reporting of K–12 program elements to support replication	2012–2016	SIGCSE TS, ICER, TOCE	92	primary and secondary education, research report or experience report, pre-college computing activities
Margulieux et al. [41] 2019	standardization of measurement in CER	2013–2017	CSE, TOCE, ICER	197	human-subjects research that measured variables excluded reviews, evaluation papers, case studies, measurement validation studies
Sanders et al. [58] 2019	inferential statistics	2005–2018	ICER	270	full papers
Decker and McGill [15] 2019	research-based evaluation instruments	2012–2016	csedreesarch.org, ICER, TOCE, CSE, Americal Evaluation Association, STELAR The Pear Institute Institute for the Integration of Technology into Teaching and Learning, MSPNet, Engineering is Elementary	47	research-based evaluation instruments available in publications or public websites and databases
Hao et al. [24] 2019	replications in CER literature	2009–2018	SIGCSE TS, ICER, ITiCSE, TOCE, CSEJTLT	54	articles including replicat[a-z]* manual review for replications
Simon [65] 2020	computing education paper classification	1996–2020	ACE	541	full papers
Malmi et al. [38] 2020	theoretical constructs and instruments for emotion, attitude, belief, and self-efficacy in computing learners	2010–2019	ICER, TOCE, CSE, LAK, Scopus search	50	computing domain containing theoretical constructs, including statistical models
Papamitsiou et al. [51] 2020	keyword analysis of CER literature	2005–2019	ITiCSE + WGR, ICER	1,274	full papers and working group reports
Simon and Sheard [67] 2020	computing education paper classification	1996–2019	ITiCSE + WGR	1,295	full papers and working group reports
This paper [2021]	quality of empirical reporting in CER	2014–2015	SIGCSE TS, ICER, ITiCSE, TOCE, CSE	427	full papers

classification scheme considers the context (e.g., subject matter or course); theme (e.g., what the paper is about like teaching technique); scope (e.g., extent of collaboration and work); and nature (e.g., distinction between research and practice) [62]. A recent classification of ITiCSE papers in 2020 found an increase in reported research over experience reports and that teaching and learning techniques remain a common theme in published work [67].

Papamitsiou et al. [51] [2020] categorized papers from 15 years of ITiCSE, ITiCSE **Working Group Reports (WGR)**, and ICER using keywords and abstracts. The main themes identified are around introductory programming, assessments, and student performance.

The categorization of topics covered in the literature provides guidance to researchers and practitioners about areas that are well studied and ripe for meta-review and gaps where additional work is needed for a better understanding of computing education. For example, "curriculum" is a common categorization in many of the above surveys, however, they each have their own nuance [16, 32, 33, 51, 62]. While our work did not identify the context of the reviewed study (e.g., CS1, databases), we do consider the subject of the evaluation (e.g., curriculum or tool), which has overlap with categorizations from related work.

2.1.2 Classification by Study Type. Several studies have classified CER literature by study type, which could include high-level categories of quantitative and qualitative or more granular categories such as experiments, quasi-experimental studies, and experience reports.

In one of the oldest CER surveys, Valentine [72] [2004] examined 444 CS1/CS2 papers from SIGCSE TS between 1984 and 2003 and found that only 21% included experimental evaluation or were experience reports. Other categories considered were Marco Polo ("I went there and I saw this"), Philosophy, Tools, and John Henry ("outrageously difficult") papers. There was an increase in the number of experimental papers presented in the past 10 years of the study period.

Randolph et al. [56] [2008] found that of 352 papers published in various CER venues between 2000 and 2005, 40% contained only anecdotal evidence and of the less than one-third of papers that did have experimental or quasi-experimental designs, 54.8% of those papers used a weaker post-test only design. Randolph et al. [56] categorized their sampled papers against Valentine [72]'s categorization and found that 40.9% of sampled papers were experimental or experience reports. Further, Randolph et al. [56] used their own categorization of research methodologies and found that of the papers that reported human subject research (n = 144), 64.6% were experimental or quasi-experimental, 26.4% were qualitative, 18.1% were causal comparative, 10.4% were co-relational, and 7.6% were survey research.

Malmi et al. [40] [2010] identified two dimensions that described the type of research: purpose and framework. They found that 86% of papers had an evaluative purpose and that 79% of papers reported a research framework. The most common research frameworks were survey (39%), experimental (15%), constructive (14%), and grounded theory (13%).

Ihantola et al. [29] [2015] conducted a survey on educational data mining research published at various venues between 2005 and 2015. Of the 76 papers that met their inclusion criteria, 78% described studies in a natural setting. Only 14% reported on formal experimental research.

Al-Zubidy et al. [5] [2016] found that 162 (70%) of the papers reviewed had some form of empirical evaluation. Their definition for empirical studies was broader than the definitions used by Valentine [72] and Randolph et al. [56]. However, the detailed evaluation type numbers show that 28% of the papers were experimental, suggesting some increase over earlier surveys.

Lishinski et al. [36] [2016] found that 71% of CSE and 87% of ICER papers between 2012 and 2015 reported empirical results and that 26% from CSE and 19% from ICER were experimental as defined by Randolph et al. [56]. The percentages of experimental work were much lower when considering the more specific definition of single group post-test only [36].

Early surveys of CER literature [56, 72] reported low rates of empirical work in the surveyed papers. More recent surveys (e.g., the past 10 years) of CER literature have shown an increase of empirical work published at a variety of CER venues. While two of the surveys reviewed literature from the CER-focused venues of ICER and CSE [36, 40], other surveys did consider other venues [5, 29] suggesting the increase in empiricism is happening more broadly in the computing community. However, with differences in the definitions of empirical work between surveys, direct comparisons cannot be made.

2.1.3 Theory. Many CER surveys discussed the use and creation of theory.

Malmi et al. [40] [2010] found that 60% of the surveyed papers reported explicit use of **theories, models, frameworks, and instruments (TMFI)** in their studies. Additionally, they found a great diversity of TMFI used during that time frame—68 distinct resources out of 78 instances.

Tenenberg's [71] [2014] position paper argues for the importance of recognizing the theories and theoretical frameworks that underlie CER research. His argument is that authors should recognize the theoretical frameworks utilized when creating research questions, because frameworks may introduce limitations on the types of questions asked and the methods used to answer research questions. Including the theoretical stance of the authors is an important aspect of reported works and serves as a foundation for future CER.

Malmi et al. [39] [2014] conducted a deeper review of CER literature that considered additional years and venues to identify the use of theories, models, and frameworks in the literature. They found that 51% of papers described at least one of 216 distinct theories, models, or frameworks. The interdisciplinary nature of CER benefits from the use of prior theoretical work, especially from outside fields. However, the disparate usage of theory, models, and frameworks is a challenge for creating a "stable theoretical base" for CER.

Lishinski et al. [36] [2016] reviewed CSE and ICER literature between 2012 and 2015 to identify the use of outside educational and learning theory and the methodological quality of the research using indicators that build on prior work (e.g., References [39, 56]). They identified an increase in the use of theory to support research and in the reporting of empirical CER work. However, they found many studies utilized less rigorous methods.

Nelson and Ko [48] [2018] summarize the use of theory in CER and identify three areas of concern in the community including tensions between explanation goals and design goals, lack of work on domain-specific theories, and publication bias due to theoretical lens of the work. They also provide concrete suggestions to move the community forward by focusing on design and using theory as a guide, investing in the creation of CER-specific theories, and reducing reviewer bias in regards to manuscripts with novel designs where theory may not be appropriate.

Malmi et al. [38] [2020] reviewed 50 papers with theoretical constructs around the topics of emotions, attitudes, beliefs, and self-efficacy of computing learners from a variety of venues between 2010 and 2019. They found that three-quarters of the papers were published between 2014 and 2019, suggesting a maturation of the CER field in using theory and validated instruments. More recent papers on theoretical constructs relied on quantitative methods rather than qualitative.

There are theoretical underpinnings to CER literature, and authors utilize work from outside of computing to support their research questions. More recent years have shown an increase in the use of theory to support the increased empirical work in CER. However, the development of CER theories is an open area of future work as the field matures. While we do not consider theory in our review of CER literature, there are enough surveys on the topic to include a discussion here for completeness.

2.1.4 Methods & Analysis. Other reviews examined the published literature to identify the type of research methods and analysis techniques used in CER studies.

Randolph et al. [56] [2008] found that of the 144 studies that utilized human participants, 107 (74.3%) used quantitative methods. Only 15.3% of the studies used qualitative methods and 10.4% of studies used mixed methods. The authors found that 44 of the 352 papers reviewed used some form of inferential statistics. Additionally, 120 of the 123 of the behavioral, quantitative, and empirical studies reported effect size.

Sheard et al. [61] [2009] conducted a review of papers on the teaching and learning of programming, published in various CER venues between 2005 and 2008, which they classified as empirical. Of the 164 papers reviewed, 37% reported quantitative results, 21% reported qualitative results, and 45% utilized mixed methods.

Malmi et al. [40] studied ICER papers published between 2005 and 2009 and found a variety of analysis methods used. These methods included statistical analysis (42%), exploratory statistical analysis (17%), descriptive statistics (11%), interpretive qualitative analysis (35%), and interpretative classification or content analysis (26%).

Al-Zubidy et al. [5] [2016] reported that 54% the papers with empirical results in SIGCSE TS 2014 and 2015 proceedings utilized surveys and 37% utilized experimental methods, which were typically quantitative. Very few papers utilized qualitative methods like observations. Margulieux et al. [41] [2019] found 32% of papers published in TOCE, CSE, and ICER between 2013 and 2017 collected both qualitative and quantitative data and over half used multiple measures.

Sanders et al. [58] reviewed the use of inferential statistics in ICER papers published between 2005 and 2018. They found that 51% of papers used inferential statistics. However, they noted that the reporting associated with inferential statistics tended to lack precise test names, confidence levels, and p-values for results. Other noted concerns with statistical test reporting included a lack of detail about data preparation, missing discussion about the assumptions for statistical tests, lack of explanation for "obscure" tests, lack of corrections for multiple statistical tests, and lack of statistical significance and effect size discussions.

The literature suggests that published CER work tends to use quantitative methods, but there is movement towards using multiple measures and appropriate statistical tests when answering research questions.

2.1.5 Participants & Context. Several reviews categorized how CER literature reports on the participants of the research study and the context of the intervention(s).

Ihantola et al. [29] [2015] reported that 34% of the educational data mining studies in their survey did not report any details about course context, which could include the course level, programming language, and topics. Additionally, 17% of the studies did not report the number of students in the study, details about student level, or demographics.

Al-Zubidy et al. [5] [2016] found that 11% of papers did not report the number of participants in their study. Additionally, when trying to identify the number of participants from the papers, there were discrepancies between reviewers due to inconsistent reporting.

McGill et al. [45] [2018] reviewed 92 papers from SIGCSE TS, ICER, and TOCE between 2012 and 2016 related to pre-college computing activities and how well papers reported key information to support future replication. One area studied was the activity component data or information about the activities and context of the intervention. They found that papers infrequently reported learning outcomes (25%), curriculum (33%), number of students in the activity (41.7%), and details about the duration of the activity within the larger context (78.6%), but many did not list the number of contact hours, which is important in K–12 work. They additionally found the reporting of both instructor and student demographic data lacking. The authors provided a checklist of recommendations to support better, and more consistent, reporting of subject and context data to support replication.

Luxton-Reilly et al. [37] [2018] found that a "large proportion" of publications related to introductory programming did not provide sufficient contextual details about the reported study, which increases the difficulty for readers to determine transferability of the results. They recommend that authors report details about the population of students and teaching context.

Margulieux et al. [41] [2019] identified several empirical reporting best practices that are lacking in CER, particularly related to the context of the study. While 85% of papers reviewed reported the number of participants, only 49% reported additional participant characteristics (e.g., prior knowledge and basic demographics). Other contextual information that would be beneficial to include are details about the task and learning environment.

Literature surveys that considered the context and participants in reviewed research found gaps or inconsistencies in reporting related to these items. Several reviewed papers excluded contextual information about the classroom or environment of the study [29, 37, 41, 45]. Several studies found that between 10%–20% of papers did not report the number of participants [5, 29, 41].

2.1.6 Data Sources. Several studies reported on the source of the data collected, including the use of validated and non-validated instruments.

Randolph et al. [56] [2008] reported on the independent, dependent, and mediating/moderating variables in 123 behavioral, quantitative, and empirical articles in their sample of CER literature. The most common independent variable was student instruction (98.9%). The most common dependent variables were attitudes (60.2%) and achievement in computer science (56.1%). Only 29% of papers used mediating or moderating variables, including gender, grade level, and learning styles. The identification of measures showed that questionnaires were the most commonly used (52.5%) followed by grades (29.3%), teacher- or researcher-made tests (22.0%), and student work (17.9%).

Sheard et al. [61] [2009] found that empirical papers from a variety of CER venues utilized formal course assessments (42%), tasks students complete (38%), and questionnaires (33%) as the top three data-gathering techniques. The authors note that the use of "established measurement instruments" was low in reviewed studies and that publications lacked details about how the instruments were used.

Malmi et al. [40] [2010] classified the data sources from ICER papers as (1) naturally occurring, (2) research specific data, (3) reflection, or (4) software. The results showed 79% of those papers used research specific data, i.e., data collected specifically for the needs of the research through interviews, questionnaires, observational data, assignments, or tasks.

Al-Zubidy et al. [5] [2016] found that most SIGCSE TS papers that used empirical evaluation reported on pedagogical techniques followed by courses and curriculum. Additionally, they found that over 75% of authors reported on new subjects specific to the paper. There was little replication or even reuse of data, even from the authors' previous studies.

Ihantola et al. [29] [2016] found in their survey of educational data mining literature that 81% of studies reported work from a single institution, 80% of studies did not consider longitudinal data, and 66% of studies only considered a single course. Additionally, they considered methods and analysis from several perspectives including details about the tasks students performed as part of the study, the type of data collected (e.g., logging, key-stroke), and the analysis methods. They found several gaps in reporting for each of these.

Decker and McGill [15] [2019] consider evaluation instruments as a key data source in CER literature. They found 47 evaluation instruments that measured cognitive, non-cognitive, and program evaluation constructs that are useful for CER researchers and would complement other quantitative and qualitative measures. They categorized the instruments based on number of items, type of items, target demographic, reliability, and validity as reported in the literature.

The sources of data for CER work covers a variety of measures from student grades, attitudinal surveys, key-logging, and other automated data collection. A variety of evaluation instruments are

available to measure various constructs [15]. Several of the literature surveys found that data was created specifically for the research study [5, 40] and focused on a single class or institution [29]. Many surveys noted that reviewed literature lacked details in reporting on the data collected and that reuse of data was low [5, 29].

2.1.7 Comparisons & Replications. Several recent papers consider the impact of reporting quality on the possibility for comparisons between and replications of CER work.

Ihantola et al. [29] [2015] identified five studies (7% of their selected papers) as replication studies in the domain of educational data mining. Al-Zubidy et al. [5] [2016] found that 46% of papers reviewed had no comparison point as part of the study. The remaining papers either had a comparison to historical data or a comparison to a dataset created specifically for the study. Additionally, fewer than six papers each year of the review (2014 and 2015) reported replications of prior work [5].

Ahadi et al. [4] [2016] surveyed 73 CER researchers about their perspectives on the value of replication work in the community. They found challenges related to the language used to describe replication/reproducibility, bias and incentives towards original work, and a lack of community value on replication work. An example of a language challenge is related to the definitions we use for replication. A survey quote suggests that replications are challenging due to the highly contextualized nature of the research environment. The authors further suggest that terminology is used inconsistently in the field [4]. However, a replication in a new environment is a *conceptual replication* and can help generalize the work [19, 59].

McGill [42] [2019] summarizes current work on replication, reproducibility, and meta-analysis in CER and provides a call to action on how to improve the field. The first suggestion is to improve individual studies, because replication, reproducibility, and meta-analysis rely on high-quality reporting. Other suggested actions are to pre-register studies, support open science, invest in large-scale collaborative research, report power analysis and effect size, create systems to store data and research tools, and incentivize replication through community support.

Margulieux et al. [41] [2019] found that while CER literature has adopted measurement instruments outside of computing and has created computing-specific instruments, most papers do not use standardized instruments to measure variables of interest. Use of standardized and validated instruments increases the reliability and validity of study results and will support meta-analysis by providing a common measure to compare across studies.

Hao et al. [24] [2019] completed a systematic literature review on replications in the CER community (including SIGCSE TS, ICER, ITiCSE, TOCE, CSEJ) between 2009 and 2018. Of the 2,269 articles, only 54 (2.38%) were considered a replication (as defined by Schmidt [59]). Of the 54 replications, 63% were successful replications. The others reported failures or mixed results. Three-quarters of the replications were conceptual (e.g., methods varied from the original study), while the remaining replications were direct (e.g., methods were as similar as possible to the original study).

CER literature does contain replications and comparisons, but replication studies are a small percentage of published CER work [24]. The CER community views direct replications as challenging, with most replication studies being conceptual replications [4, 19, 24, 59]. Comparisons and replications can be supported with the use of validated instruments [15, 41] and other open science practices can support replications [42].

2.1.8 Gaps in Reporting. Many literature reviews included a discussion about gaps in reporting of the literature surveyed. These gaps are places where the literature review was more challenging because papers were missing key pieces of information.

Randolph et al. [56] [2008] reviewed 352 papers randomly selected from all papers published in several computing education venues between 2000 and 2005. As part of their methodological review, they identified whether papers contained elements considered important by the American Psychological Association [7], which include items such as an abstract, research questions, and a description of participants. They found that less than 50% of the sampled papers reported purpose/rationale (36.6%), research questions/hypotheses (22.0%), participants description (45.5%), procedures (37.4%), and separate results and discussion (29.3%).

Sheard et al. [61] [2009] report on a lack of connection between findings related to the process of learning programming and relevant theories and models of learning. Malmi et al. [40] [2010] found it challenging to identify the theories, models, frameworks, and instruments utilized in the papers reviewed with similar challenges in other dimensions considered during the review.

Ihantola et al. [29] demonstrated the challenges to re-analysis, replication, and reproduction resulting from a lack of context and details in existing literature through three case-studies. Additionally, their analysis of quality measures on reviewed papers found that reporting related to confounding factors, threats to validity, and ethical issues were lacking. Some survey responses from Ahadi et al. [4]'s [2016] work suggests that replications can be difficult due to gaps in the reporting of a research study.

Al-Zubidy et al. [5] [2016] identified several areas where gaps in reporting lead to challenges when reviewing the literature particularly with regard to lack of replication and inconsistent paper organization. In particular, they found that between 40% to 70% of papers in the years evaluated from the SIGCSE TS were lacking threats to validity.

Lishinski et al. [36] [2016] found that 47% of CSE and 56% of ICER papers during their study timeframe reported explicit research questions. While this level of reporting is an improvement over the use of research questions as reported by Randolph et al. [56], the lack of explicit research questions in some empirical studies is a concern. Lishinski et al. [36]'s finding about research questions are lower than the 79% reported by Ihantola et al. [29] possibly due to the focus on educational data mining and learning analytics.

McGill et al. [45] [2018] found many gaps in reporting of pre-college activities related to the study context and the demographics of instructors and students. They provide a checklist with recommendations that would also be useful for education researcher more broadly.

Luxton-Reilly et al. [37] [2018] found in their review of introductory programming that many papers lacked details about the construct/intervention and its operationalization and effect sizes, which limits replication. They suggest archiving course information including "syllabus, learning outcomes, infrastructure, teaching approaches, and population demographics...." Prior work found several gaps in reporting quality [5, 29, 36, 56], lack of connection between results and theories [40, 61], and lack of contextual information [29, 37, 45]. Open science practices, like archiving research study information, could supplement the details in the literature [37, 42].

2.2 Guidelines for Reporting and Quality

Several organizations provide guidelines for reporting and assessing study quality. We encourage interested readers to review these resources to consider additional quality expectations when designing and reporting on empirical CER studies.

The **What Works Clearinghouse** (WWC) has developed a *Standards Handbook* [12] that provides resources for a systematic review process, including the development of a quality rubric, to provide summary reports about empirical educational research that can be of use to policymakers. The structured review is intended to assess the *internal validity* of the study. WWC focuses on four types of research: randomized controlled trials, quasi-experimental design, regression discontinuity design, and single-case design. Studies related to a particular area of interest and quality

rubric provide one of three labels: meets WWC Design Standards without reservations, meets WWC Design Standards with reservations, does not meet WWC Design Standards. For example, quasi-experimental designs that lack randomization can only achieve the category "meets WWC Design Standards with reservations" due to the lack of randomization in the process. The WWC Design Standards provide a deeper quality assessment, specifically around the methods, results, and internal validity of educational research studies than our rubric on reporting norms.

The CONSORT 2010 Statement [60] provides guidelines and a checklist for reporting parallel group randomized trials. While the authors of CONSORT do not prescribe a specific article structure, they do suggest using subheadings to help readers find information in the manuscript. Appropriate subheadings provide readers with guideposts on where to find important information and can be especially useful for supporting possible replications.

The **American Psychological Association (APA)** provides *Journal Article Reporting Standards (JARS)* for researchers, reviewers, and editors [7]. The website provides guidelines for reporting standards on quantitative, qualitative, and mixed research publications and study designs. It also provides guidance on meta-analysis reporting standards.

The **American Education Research Association (AERA)** *Standards for Reporting on Empirical Social Science Research in AERA Publications* provides guidance for reporting on empirical educational research [6]. Guidance includes a clear statement of the problem, description of the study design, overview of the data collection and sources of evidence, a description of key measurements and how they are operationalized by the study, details of analysis procedures, scope, and ethics.

Within the computing education community, several articles provide guidance on high-quality reporting of CER work. Daniels and Pears [14] provide a framework for designing action research to support answering research questions about "concrete teaching and learning challenges" in computer science classrooms. Their framework provides guidance on reporting the researcher's epistemology and theoretical perspective(s), describing the context of the research study, and describing the methodology and methods associated with the research question. McGill and Decker [43] describe the challenges in creating a repository for K–12 CER research due to inconsistencies in reporting standards and questions about study quality identified by a focus group gathered to create repository requirements. Some of the findings from the focus group suggest that reporting on basic technical components would support replication. The www.csedresearch.org website serves as a repository for instruments and guidelines for reporting on K–12 studies, which builds on other work by McGill, Decker, and others [45]. As part of their systematic literature review, Ihantola et al. [29] created a quality assessment rubric for selected studies that reported on educational data mining and learning analytics.

Researchers in related fields that also consider human subjects, like software engineering, have proposed reporting guidelines for empirical work. Runeson and Höst [57] describe guidelines for reporting case studies, including suggested headings and subsections. Reference [57] compares case study reporting guidelines to experimental reporting guidelines created by Jedlitschka and Pfahl [31] and refined by Kitchenham et al. [35]. Additionally, Kitchenham et al. [35] provide checklists that researchers, practitioners, meta-reviewers, replicators, and reviewers should consider when reading experimental work. Carver [10] provides guidelines for reporting experimental replications in software engineering that can also be useful as general reporting guidelines for empirical work. As part of reporting on a replication, the original study should be discussed, including the research question(s), participants and their characteristics and context, experimental design, artifacts or resources used in the study, context variables, and summary of the results as general reporting guidelines for empirical work.

2.3 Guidelines for Replication

Our focus on reporting quality is in the goal of supporting replications, meta-analysis, and theory-building in the CER community. This section describes key references on replications in CER and more broadly.

Schmidt [59] provides a pragmatic and operational overview of replication, particularly in the social sciences. Their definitions of direct replication and conceptual replication form the foundation of the discussion of replication and capture our views of replication. Replications provide additional confirmatory power about a research question. Schmidt suggests five functions of replication: "to control for sampling error, to control for artifacts, to control for fraud, to generalize results to a larger or to a different population, and to verify the underlying hypothesis of the earlier experiment." While Schmidt [59] provides a framework for replications and reproductions, he additionally provides a pragmatic reflection on practical publications of reflections given community biases towards original work by using *follow-up studies* and *systematic replications*. Follow-up studies may include direct replication of earlier work and additional experimental conditions that either provide generalizability or novelty beyond the replicative piece. Systematic replications provide a way of exploring the variation around a research replication modeled as a data matrix. Replications consider different cases within the matrix to systematically explore the research space.

The **National Science Foundation (NSF)** and the **Institute of Education Sciences (IES)** issued joint *Companion Guidelines on Replication & Reproducibility in Education Research* [19] as a supplement to the *Common Guidelines for Education Research and Development* [18]. The *Companion Guidelines* define reproducibility and replication and describe the importance of both in furthering theory in education. The guidelines for designing reproducible studies suggest that "Analyses should be described in sufficient detail as to allow other researchers to reproduce the results using the same dataset." When reporting results, researchers should share data and analysis details, how results compare to replicated or reproduced studies, and specific details about data that was excluded or omitted [19]. The definitions from the *Companion Guidelines*, which are similar to those provided by Schmidt [59], are:

- *reproducibility*: "the ability to achieve the same findings as another investigator using extant data from a prior study" [19].
- *replication*: "involved collecting and analyzing data to determine if the new studies (in whole or in part) yield the same findings as a previous study" [19]. Replication studies are further broken down into two categories:
  - *direct replication*: "seek to replicate findings from a previous study using the same, or as similar as possible, research methods and procedures as a previous study" [19, 59].
  - *conceptual replication*: "seek to determine whether similar results are found when certain aspects of a previous study's method and/or procedures are systematically varied" [19, 59].

Nosek and Lakens [49] describe the benefits of direct replication as increasing the size of the data that can help identify false positive results, help establish generalizability of results, and identify the boundaries of results.

Gómez et al. [21] explore the language and definitions of how other disciplines verify findings of experimental research. A review of literature found 18 sources with classifications of replications. Their synthesis identified three key groupings of definitions based on whether the authors use the **same methods** for either *operational replication* of the methods or *empirical generalization* of the results, **different methods** for a *conceptual replication* to determine if the results around a given hypothesis are reproducible, or **existing datasets** where the analysis is replicated for *internal replication* or for a *reanalysis of the data* using different methods. Gómez et al. [21] provide the following terms and definitions:

- *Re-analysis*: uses the same or different analysis on "the data of a previously run experiment are used to verify the results rather than re-running the experiment" [21].
- *Replication*: uses the same methods with different data that "verifies that the observed findings are stable enough to be discovered more than once" [21].
- *Reproduction*: uses the same hypothesis, but different methods and data to verify "that the findings are not to be attributed to the experimental method" [21].

Ihantola et al. [29] build on Gómez et al.'s [21] work to create a novel classification of reproduction studies in the R.A.P. Taxonomy. They consider three key criteria: *researchers*, *data analysis*, and *data production* to identify seven classifications where one or more of the criteria are changed for study reproduction. By considering the taxonomy, researchers can examine the state space of possible study reproductions *related to a specific hypothesis*. The seven categories and their definitions are:

- *Re-Analysis*: "a different different experimenter is following the same analysis done before with the original data set for review purposes" [29].
- *Extended Analysis*: "an experimenter is extending the baseline study by looking at a previously analyzed data set, but using new analysis methods" [29].
- *Repetition*: "an experimenter is repeating the same analysis with new data" [29].
- *Verification*: "the same data set is looked into again by a different experimenter and using a different analysis method to verify the conclusions" [29].
- *Replication Study*: "a different experimenter is following the same analysis method as in the baseline study, but using their own different data set" [29].
- *Triangulation*: "an experimenter is collecting a new data set to be analyzed with a new method" [29].
- *Reproduction*: "a different experimenter is analyzing their own new data set and following a new analysis method designed for the study in order to test the hypothesis in the baseline study" [29].

ACM updated their Artifact Review and Badging definitions [3] in 2020 to correspond to definitions used by the **National Information Standards Organization (NISO)**. ACM uses the following definitions:

- *Repeatability*: "Same team, same experimental setup" [3].
- *Reproducibility*: "Different team, same experimental setup" [3].
- *Replicability*: "Different team, different experimental setup" [3].

The terminology utilized by NSF and ISE *Companion Guidelines* [19], Gómez et al. [21], Ihantola et al. [29], and ACM [3] is inconsistent. Table 2 maps the definitions to each other. We utilize the definitions from the *Companion Guidelines* [19] in our discussion.

2.4 Contributions

The recent increase in characterizing the reporting standards in CER literature through various reporting lenses demonstrates the need for the community to identify and use common reporting standards. We see our contribution as complementing the recent work of others [4, 5, 24, 29, 36, 38, 41, 42, 44, 45, 48, 58, 71] to improve reporting standards in the CER community. Specifically, the current article provides an advance over prior work in three dimensions:

- *Scale* - we considered the proceedings of five CER venues for a full two years, including the SIGCSE TS;

Table 2. Alignment of Replication Definitions

NSF/ISE Companion Guidelines [19]	Gómez et al. [21]	Ihantola et al. [29]	ACM [3]
Reproducibility	Re-analysis	Re-analysis Verification	Repeatability
Direct Replication	Replication	Repetition Replication Study	Reproducibility
Conceptual Replication	Reproduction	Extended Analysis Triangulation Reproduction	Replicability

- *Scope* - our study provides a broad snapshot that covers the full proceedings/issues of CER venues compared with the more focused snapshots of prior work.
- *Focus* - our study investigates the quality of empirical reporting in CER rather than classification.

3 METHODOLOGY

Our goal is to describe the reporting of empiricism across the most common venues for computing education research and practice. We have identified elements of empiricism that should be included in reports of research results, which parallel the key items needed during design of an educational research study [8, 13, 16]. Our systematic literature review [34] will reveal the state of the practice for reporting empirical research results and provide guidance to the CER community for how to improve reporting to provide the basis the community needs to advance in a scientifically rigorous manner.

Our study investigates the following **research questions (RQs)**:

- RQ1: What percentage of papers in CER venues have some form of empirical evaluation?
- RQ2: Of the papers that have empirical evaluation, what are the characteristics of the empirical evaluation?
- RQ3: Of the papers that have empirical evaluation, do they follow norms (both for inclusion and for labeling of information needed for replication, meta-analysis, and, eventually, theory-building) for reporting empirical work?

3.1 Selection Criteria

Previous work done through 2008 [56, 72] showed a small increase in the presence of empirical evaluation in papers in CER conferences in general. Several more recent papers show an increase in empirical work across a variety of venues [5, 29, 36, 40]. Our evaluation includes two years of papers from five venues (SIGCSE TS, ICER, ITiCSE, TOCE, and CSE). We included all papers from the TOCE and CSE editions and two iterations of each conference during 2014 and 2015.

3.2 CER Empiricism Assessment Rubric

We developed and piloted the first version of the *CER Empiricism Assessment Rubric* in 2015 [5]. This initial version of the rubric was based on the evaluation rubrics used in previous reviews of CER literature [56, 72] and on items considered when designing empirical studies [8, 16]. Specifically, the rubric was focused on aspects of a research paper that are essential for others to understand and potentially replicate the experiment, as evidenced by prior work [56, 72] and supported by guidelines in educational research [12, 19, 59, 60]. Randolph et al.'s rubric, however, incorporated more granular details in its categories. For example, Randolph et al.'s rubric took

into account the computing domain from which the research intervention originated, including categories such as "Visualization" and "Simulation" as discrete options [56]. Our rubric takes a more high-level approach, focusing more on the general role of the intervention in how it relates to computing education. In this example, instead of noting whether the intervention incorporated "Visualization" or "Simulation," we focused on whether the intervention was an assignment, a tool, a curricular innovation, and so on.

After we applied the initial version of the rubric in our study of the SIGCSE TS proceedings, we evolved both the rubric and the methodology for applying it to a large number of papers to arrive at the current version of the rubric found in Appendix 7. In updating our rubric, we referenced the evolving reviewing guidelines for ACM SIGCSE conferences and work in the **scholarship of teaching and learning (SoTL)** space to ensure that our core rubric items were capturing current best practices in CER work [8]. Further, we streamlined our evaluation methodology to make the application of the rubric to a large set of papers more feasible, while still maintaining the quality of our evaluation by examining the inter-rater reliability. Finally, we discovered some papers that exposed corner cases that we needed to address, such as papers that purported to be empirical studies, but reported no number of actual participants in the study.

The rubric is composed of a *Base Rubric* that is applicable to all papers with some level of empirical results and three additional rubrics for specific categories of research projects.

The *Base Rubric* contains a set of characteristics that captures the overall nature of the empirical work including information such as the type of evaluation method, the evaluation subject, whether there is any comparison, and the number of participants. These items provide a high-level characterization of the empirical work presented, answering questions such as:

- What is the balance between quantitative and qualitative work?
- At what rate do researchers publish studies on curricula?
- Are researchers performing replication studies or creating their own interventions?
- How many participants are in the typical CER study?

Depending upon the type of empirical study included in the paper, i.e., quantitative, qualitative, survey, and/or descriptive, we then employed one or more additional rubrics. Each of these rubrics allowed us to characterize the presence and clarity of important information that CER papers should include relative to each type of study. The items on these additional rubrics all use the same categorization scale that identifies whether a piece of information is *present* and how easy it is for the reader to *identify* that piece of information. The scale captures both types of information as follows:

- *Completeness - the level of completeness of presented information*
  - Complete - Answers/addresses all questions for a rubric category. There is no assessment on the quality of the answer.
  - Partial - Answers/addresses some of the questions for a rubric category. There is no assessment on the quality of the answer.
  - Not Present - Answers/addresses none of the questions for a rubric category. The questions should be addressed.
  - Not Applicable - The rubric item is not applicable to the paper.
- *Labeled - whether the presented information is clearly labeled*
  - Labeled - There is a heading appropriate for the rubric item or there is emphasis (bold/italics) for the rubric item.
  - Not Labeled - There is no heading or emphasis for the rubric item to easily find the item in the paper.

Therefore each rubric item can take one of six labels:

- Complete and Labeled
- Complete and Not Labeled
- Partial and Labeled
- Partial and Not Labeled
- Not Present
- Not Applicable

The items included on the rubric represent the type of information necessary to support reproducibility and replication of research studies. We use the definitions from the *Companion Guidelines on Replication & Reproducibility in Education Research* published jointly by the **National Science Foundation (NSF)** and the **Institute of Education Sciences (IES)** [19].

- *reproducibility*: "the ability to achieve the same findings as another investigator using extant data from a prior study" [19].
- *replication*: "involved collecting and analyzing data to determine if the new studies (in whole or in part) yield the same findings as a previous study" [19]. Replication studies are further broken down into two categories:
  — *direct replication*: "seek to replicate findings from a previous study using the same, or as similar as possible, research methods and procedures as a previous study" [19, 59].
  — *conceptual replication*: "seek to determine whether similar results are found when certain aspects of a previous study's method and/or procedures are systematically varied" [19, 59].
- *meta-analysis* - is a quantitative, formal, systematic analysis of a number of related study results to identify general patterns and draw overarching conclusions about the entire body of research [23].
- *theory-building* - is the process of using replication, repetition, and meta-analysis to systematize knowledge. Theories can drive the generation of hypotheses and produce predictive theory through scientific enquiry [16]. Additionally, conceptual replications can produce understanding and confirm underlying theory [59].

For the remainder of this article, we use the term *replication* to cover any type of study design that builds on prior study designs through replication in the same setting (i.e., direct replication) or through systematic variance (i.e., conceptual replication).

Reporting common information and using open science supports replications [19]. Meta-analysis can then utilize these common results, aggregate them, and increase overall generalizability of findings to "gain a better understanding of what interventions improve (or do not improve) educational outcomes, for whom, and under what conditions" [19]. Because individual studies are more prone to error or bias, meta-analysis or meta-studies provide the opportunity to synthesize results around a research question to understand impacts on a broader set of learners [5, 19, 42, 59]. This process provides for a mechanism for theory, either confirmatory of educational theory in a computing context or the emergence of theory to support new phenomenon [16, 42, 59].

For example, good practice is for a paper to provide an overview of the participants, including demographics and the sampling or recruitment method [45]. If a paper met all the criteria specified in the rubric, then the paper receives a label of "Complete" for that rubric item. If the paper meets at least one, but not all, of the criteria, then the paper receives a label of "Partial" for that rubric item. If the paper meets none of the criteria in the rubric, then the paper receives a label of "Not Present" for the rubric item. Finally, if this particular rubric item does not apply to the paper, then the paper receives the label of "Not Applicable" for that rubric item.

To ease the replication, meta-analysis, or theory-building process, it is important for a paper to not only contain the appropriate information but also to present that information in a way that readers can easily locate it, similar to the checklists in CONSORT [60] and suggested information for reporting in APA JARS [7]. We determined a rubric item to be easy to find if there was a relevant section heading in the paper for the item. We also considered text that emphasized a specific item through italics, bolding, or a bulleted list. These are common ways of noting research questions rather than using a section header. We assigned a value of either "Labeled" or "Not Labeled" for each rubric item in each paper.

To reiterate, the purpose of the *CER Empiricism Assessment Rubric* is to characterize the study and the rigor in which a paper reports an empirical study. The goal is to comment on the efficacy of the reporting from the perspective of a reader who wishes to replicate, perform meta-analysis, or build theories. The rubric does not comment in any way on the quality of the study or reporting itself, beyond the presence of the expected information.

We completed the *Experimental Rubric* for each paper that contained some form of empirical evaluation. This rubric characterizes the rigor in which the paper reports and labels the key aspects of an empirical study [8, 13, 16, 60]. These aspects include concepts and questions such as:

- Are the research objectives obvious and easy to find?
- Do the authors present related work?
- Is the study design properly presented?
- How was the data gathered and analyzed for this work?
- Are the results presented in a succinct, direct way?
- Are threats to validity discussed?

For an empirical study to be replicated, a paper needs to thoroughly discuss and clearly label these items so readers can easily find them.

We completed the *Survey Rubric* for each paper that uses a survey as either the primary research methodology or one of the data sources. Within those topics, the rubric items examine whether the paper describes the survey creation process, the rationale behind the questions, and the execution of the survey, including its administration and the medium used.

We completed the *Descriptive/Persuasive Rubric* for papers that do not claim any cause/effect relationship. These papers are often presented as position papers. This rubric contains items focused on the goal of the paper's argument, the presence of related work, the soundness of the argument, and whether supporting evidence is present.

3.3 Gathering Data and Applying the Rubric

Because we included all papers from each venue in the chosen years, we manually extracted the DOI entries for each paper from the digital library. Then, we downloaded the PDFs of the papers to analyze locally.

In an initial phase, we examined the first 50 papers to refine our methodology. We randomly assigned each paper to two researchers for initial categorization. Each researcher independently evaluated the paper using the rubric. Then, we met to discuss any discrepancies in the analysis. If needed, then we asked a third researcher to review a paper to resolve any discrepancies. At the end of this process, all researchers agreed on the final categorization of the paper.

Based on the experience with the first 50 papers, we modified our methodology for the remaining papers. Overall, we found very few discrepancies in this first set of papers. We computed kappa to measure the inter-rater reliability for each of the rubric items across the 50 papers. Overall, our level of agreement was very high, with the kappa values for each rubric item above 0.8 ($p < 0.01$). Based on this high level of agreement, we decided it was not necessary for two reviewers to

Table 3. Empirical Evaluation by Venue and Year

Venue	Year	# Papers	Acceptance Rate	# Empirical Papers	% Empirical Papers	# Empirical Papers by Venue
SIGCSE TS	2014	110	39.4%	91	82.7%	176
SIGCSE TS	2015	103	36.0%	85	82.5%	
ICER	2014	17	25.0%	14	82.4%	38
ICER	2015	25	26.0%	24	96.0%	
ITiCSE	2014	51	35.0%	43	84.3%	85
ITiCSE	2015	52	43.5%	42	80.8%	
TOCE	2014	21	n/a	15	71.4%	29
TOCE	2015	21	n/a	14	66.7%	
CSE	2014	11	n/a	10	90.9%	23
CSE	2015	16	n/a	13	81.3%	
Totals		**427**		**351**	**82.2%**	

complete the full rubric on each paper. However, we also did not want to fully rely on only one reviewer for each paper. As a compromise, because the *Base Rubric* contains the information that determines the type of research present in the paper and determines which other rubric(s) apply, we had two researchers complete the *Base Rubric* for each paper. After resolving any discrepancies, this process resulted in a consensus on the type of research present in each paper. Based on the type of research present, one researcher then reviewed the paper using one or more of the remaining rubrics (*Experimental, Survey, and Descriptive/Persuasive Rubrics*). Finally, we merged all results into a single spreadsheet for analysis and calculations [27].

4 RESULTS

This section describes the results of our analysis of our dataset [27], organized around the three research questions.

4.1 RQ1: Empirical Evaluation

> **RQ1: What percentage of papers in CER venues have some form of empirical evaluation?**

We classify a paper as *empirical* if its **Evaluation Method** is *Descriptive*, *Survey*, *Qualitative*, or *Quantitative* such that it contains *empirical evidence* addressing a research goal, question, or hypothesis. As shown in Table 3, 351 of the 427 papers (82%) contained some type of empirical evaluation. While this number is quite high, note that we did not assess the quality or correctness of the empirical evaluation, just the presence of it. When considering specific venues, we found that TOCE had a lower percentage of empirical papers. This result is likely due to the fact that a portion of the TOCE papers were editorials and opinion papers, which we would not expect to contain empirical evaluation. On the other end of the spectrum, 24 out of 25 papers in ICER 2015 had some sort of empirical evaluation.

When comparing these results with those from earlier literature reviews, we see an increase in the percentage of papers containing empirical evaluation over time. Valentine [72] found 21% of papers discussing CS1 and CS2 topics had "experimental" evaluation, and Randolph et al. [56] found 35% of papers with behavioral, quantitative, or empirical research from a broader set of

Table 4. Evaluation Methods by Venue

Evaluation Method	SIGCSE TS		ICER		ITiCSE		TOCE		CSE		TOTAL	
	#	%	#	%	#	%	#	%	#	%	#	%
Descriptive/ Persuasive	7	4%	0	0%	1	1%	2	7%	3	13%	13	4%
Survey	32	18%	4	11%	12	14%	7	24%	4	17%	59	17%
Qualitative	15	9%	12	32%	13	15%	4	14%	5	22%	49	14%
Quantitative	126	72%	22	58%	59	69%	17	59%	12	52%	236	67%

CS education venues from 2000–2005 contained empirical evaluation. Early surveys of ICER papers between 2005–2009 found 86% of papers had an evaluative purpose, appropriate for a conference focused on CER [40]. Lishinski et al. [36]'s analysis considered ICER and CSE papers between 2012 and 2015 and found that 71% of CSE papers and 87% of ICER papers reported empirical results. These numbers are lower than some of our results but consider a larger time frame. A review of SIGCSE TS papers from 2014 and 2015 found that 70% of papers had some form of empirical evaluation [5]. Our definition of empirical evaluation is more broad, resulting in higher numbers during the same time period. When looking at data mining and automated learning research across venues, Ihantola et al. [29] found that 78% of papers reported on a study in a natural (e.g., empirical) setting. Our findings, along with complementary work, suggest an increase in the amount of empirical work since the early surveys by Valentine [72] and Randolph et al. [56], which in turn should lead to an increase of CER in the SIGCSE community. The remaining research questions investigate the state-of-the-practice for reporting empirical studies to further understand the concerns raised in our previous work [5].

4.2   RQ2: Characteristics of Empirical Evaluation

> **RQ2: Of the papers that have empirical evaluation, what are the characteristics of empirical evaluation?**

To answer this question, we characterize the papers using the *CER Empiricism Assessment Rubric – Base Rubric*. For each characteristic, we provide a table that reports the raw number of papers with that characteristic type followed by the percentage of the total empirical papers for the venue (based on the last column of Table 3). We gave some papers multiple values for a characteristic. In those cases, the percentages may total more than 100%. Additionally, we summarize the characteristic items in the total column as the percentage of all 351 empirical papers.

Evaluation Method (Table 4). The *CER Empiricism Assessment Rubric* describes each evaluation method in detail. However, for clarity, we highlight the *Survey* and *Descriptive/Persuasive* evaluation methods more carefully here to describe their specific usage as an evaluation method. We coded papers as *Survey* when they described a community survey with the goal of describing the current state of that community (i.e., the paper does not involve any interventions). *Qualitative* and *Quantitative* evaluation methods may have surveys as one of their *Data Sources* in the study. We coded papers as *Descriptive/Persuasive* if its focus was to describe a current situation or to persuade the reader about a position. These papers do not test relationships among variables (statistically or otherwise).

Table 5. Evaluation Subject by Venue

Evaluation Subject	SIGCSE TS		ICER		ITiCSE		TOCE		CSE		TOTAL	
	#	%	#	%	#	%	#	%	#	%	#	%
Assignment	13	7%	4	11%	2	2%	2	7%	0	0%	21	6%
Community	24	14%	7	18%	15	18%	8	28%	2	9%	56	16%
Curriculum	50	28%	3	8%	13	15%	7	24%	8	35%	81	23%
Pedagogical Technique	59	34%	17	45%	34	40%	9	31%	7	30%	126	36%
Tool	24	14%	3	8%	17	20%	2	7%	2	9%	48	14%
Other	6	3%	4	11%	4	5%	2	7%	4	17%	20	6%

Table 6. Evaluation Subject Source by Venue

Evaluation Subject Source	SIGCSE TS		ICER		ITiCSE		TOCE		CSE		TOTAL	
	#	%	#	%	#	%	#	%	#	%	#	%
Authors Here	101	57%	21	55%	51	60%	17	59%	18	82%	208	59%
Authors Elsewhere	11	6%	4	11%	9	11%	3	10%	1	5%	28	8%
Other Modified	32	18%	6	16%	8	9%	1	3%	2	9%	49	14%
Other Not Modified	10	6%	7	18%	6	7%	4	14%	0	0%	27	8%
Community	22	13%	1	3%	11	13%	4	14%	1	5%	39	11%

Examining the venues in more detail reveals that the papers published at SIGCSE TS are over-whelmingly *Quantitative* followed by *Survey*. SIGCSE TS has very few *Qualitative* papers in the years studied. The ITiCSE papers are also mostly *Quantitative*, but do have a higher percentage of *Qualitative* papers. ICER and CSE had the most balanced split between *Quantitative* and *Qualitative* studies, but a smaller percentage of *Survey* papers.

Over two-thirds of empirical studies in CER literature are *Quantitative* (67%). We coded only five papers with multiple evaluation methods. These papers were either *Survey – Qualitative* or *Survey – Quantitative*. While the large emphasis that these CER papers place on *Quantitative* methods provides interesting numerical results, it does suggest the lack of nuance that *Qualitative* methods bring to help understand the *why* and *how* of a study, which can *expand* and *deepen* our under-standing of computing phenomena [8, 26].

Evaluation Subjects (Table 5). *Pedagogical Techniques*, or teaching methods, were the most com-mon evaluation subject in all venues. *Tools* papers were more common at the SIGCSE TS and ITiCSE. All venues had papers evaluating the broader *Community*. *Curriculum* papers were more common at the SIGCSE TS, TOCE, and CSE. Readers may find the distribution of paper types at these venues for 2014 and 2015 interesting, as they consider future studies that either fit with or fill gaps with existing literature.

Evaluation Subject Source (Table 6). This rubric item is a measure of replication and repro-ducibility in the community. This item could take one of five values: (1) *Authors Here* – the authors created the evaluation subject for use in the current study; (2) *Authors Elsewhere* – the authors cre-ated and published the evaluation subject in a previous publication; (3) *Other Modified* – someone other than the authors created the evaluation subject, but the authors modified it; (4) *Other Not Modified* – someone other than the authors created the evaluation subject, and the authors did not

Table 7. Comparison in Evaluation by Venue

Comparison in Evaluation	SIGCSE TS		ICER		ITiCSE		TOCE		CSE		TOTAL	
	#	%	#	%	#	%	#	%	#	%	#	%
Comparison	87	49%	20	53%	46	54%	15	52%	8	35%	176	50%
None	89	51%	18	47%	39	46%	14	48%	15	65%	175	50%

Table 8. Number of Participants by Venue

Number of Participants	SIGCSE TS		ICER		ITiCSE		TOCE		CSE		TOTAL	
	#	%	#	%	#	%	#	%	#	%	#	%
1	6	3%	0	0%	2	2%	1	3%	0	0%	9	3%
2	1	1%	2	5%	2	2%	1	3%	0	0%	6	2%
3−10	10	6%	2	5%	6	7%	3	10%	0	0%	20	6%
11−30	38	22%	7	18%	21	25%	3	10%	2	10%	71	20%
31−75	38	22%	10	26%	15	18%	6	21%	6	29%	75	21%
76−150	25	14%	6	16%	13	15%	3	10%	1	5%	48	14%
151−300	20	11%	5	13%	7	8%	3	10%	5	24%	40	11%
301−999	19	11%	3	8%	12	14%	4	14%	5	24%	43	12%
1,000+	13	7%	3	8%	7	8%	4	14%	1	5%	21	8%
Missing	6	3%	0	0%	0	0%	1	3%	1	5%	8	2%

modify it; or (5) *Community* – self-identifying group of people related to an area of interest (if the evaluation subject is *Community*, then the source is *Community* as well).

More than half of all empirical papers used evaluation subjects developed for the specific study (i.e., *Authors Here*), indicating that researchers may be working in isolation and not using existing datasets or possible comparisons to similar participants in a different context. Papers that reuse an evaluation subject from a previous study by another author demonstrate replication and sharing within the broader community [19, 59]. In some cases, the authors conducted a study on an evaluation source that they utilized in earlier work, which we coded as *Authors Elsewhere*. Overall, ICER had the highest percentage of papers in which authors utilized evaluation subjects from other researchers (e.g., *Other Modified* and *Other Not Modified*).

Comparison (Table 7). By comparing the results from an intervention to the results from a baseline approach, a paper can provide additional support to strengthen a conclusion that the intervention caused the observed effect. The lack of *Comparison* leaves open the possibility of confounding factors influencing the observed result. However, when papers report a case study (i.e., an in-depth analysis of a single case), the lack of comparison is expected (e.g., *None*). Case studies can provide valuable findings that other researchers can attempt to replicate in their own environments, using more experimental approaches. Approximately half of the papers in each venue did provide some type of *Comparison* between two or more groups of participants. To promote comparisons between researchers, Margulieux et al. [41] suggest utilizing common, standardized, measurements.

Participants (Table 8 & Table 9). The number of *Participants* in published studies can provide some insight into the strength of the conclusions drawn. Of concern are the papers that did not provide any numbers when discussing the participants in the study. Of the 349 empirical papers

Table 9. Summary Statistics of Participants by Venue

Summary Statistics	SIGCSE TS	ICER	ITiCSE	TOCE	CSE	TOTAL
Average	1,937.6	913.1	921.2	583.8	314.5	1,363.7
Median	63.5	75	60	95.5	183	72
Mode	1	75	23	5	183	12

Table 10. Type of Study for Non-descriptive Papers by Venue and for *Quantitative* Only Papers by Venue

	SIGCSE TS		ICER		ITiCSE		TOCE		CSE		TOTAL	
Type of Study (Non-descriptive)	#	%	#	%	#	%	#	%	#	%	#	%
Observational	86	51%	20	53%	37	44%	16	59%	12	60%	171	51%
Interventional	83	49%	18	47%	47	56%	11	41%	8	40%	167	49%
Type of Study (Quantitative Only)	#	%	#	%	#	%	#	%	#	%	#	%
Observational	63	50%	9	41%	20	34%	8	47%	5	42%	105	44%
Interventional	63	50%	13	59%	39	66%	9	53%	7	58%	131	56%

that should report subject data,[3] 98% of papers reported subjects in some form. This percentage is higher than the 83.7% of articles reporting subjects in McGill et al. [45]. We additionally report summary statistics for the number of participants by venue and across all venues. Since we are considering all publications in the venues rather than a subset based on an inclusion criteria, the mean and median number of participants is higher than the mean of 328 and median of 45 reported by McGill et al., likely due to the K–12 focus [45].

Type of Study (Table 10). An *Observational* study is performed in a natural setting in which the researcher collects data via observation without manipulation of the situation. Conversely, an *Interventional* study is performed by assigning participants into groups (e.g., control and experimental) and applying the treatment to the experimental group to measure its effect. Our use of observational and interventional are not intended as a direct relationship to qualitative and quantitative methods, but are instead a way of describing the setting. *Observational* studies are similar to Fincher's and Petre's [16] definition on *in situ*—the normal or natural setting. *Interventional* studies correspond to Fincher's and Petre's [16] settings of *under constraints* and *in a laboratory* where there is some direct intervention to the environment.

There is a fairly even split between the two types of studies across all venues. This result indicates a nice balance in CER work. However, when looking at *Quantitative* work only, we do see an increase in the proportion of *Interventional* studies. These results suggest that there is a gap in using *Qualitative* methods to assess CER interventions.

Data Source (Table 11). We identified seven key data sources and included an option for data sources that do not fit any of the existing categories. Overall, the most common data sources utilized in *Quantitative*, *Qualitative*, and *Survey* work were *Surveys*, *Assessment Data*, and *Automated Data*. In addition, the *Qualitative* studies frequently featured *Interviews*, *Focus Groups*, and *Observations*. Our categorization of data sources has some overlap with related work, but our focus was

---

[3]Two CSE papers categorized as descriptive had N/A recorded for subjects.

Table 11. Data Source of Papers by Venue

Data Source		Assessment Data	Automated	Course Evaluations	Interviews	Focus Group	Observations	Survey	Other	More than One	Survey ONLY
SIGCSE	Quant	46%	13%	7%	6%	1%	8%	35%	5%	19%	24%
	Qual	40%	0%	0%	47%	13%	7%	27%	0%	33%	7%
	Sur	6%	0%	0%	3%	3%	0%	100%	0%	13%	88%
ICER	Quant	45%	27%	0%	0%	0%	5%	32%	14%	23%	14%
	Qual	8%	8%	0%	67%	0%	33%	0%	17%	25%	0%
	Sur	0%	0%	0%	0%	25%	0%	100%	0%	25%	75%
ITiCSE	Quant	53%	17%	2%	5%	0%	12%	47%	10%	41%	15%
	Qual	23%	15%	0%	15%	8%	15%	15%	15%	8%	8%
	Sur	0%	0%	0%	0%	0%	0%	100%	0%	0%	100%
TOCE	Quant	47%	12%	0%	12%	6%	12%	29%	6%	24%	18%
	Qual	0%	0%	0%	50%	0%	25%	0%	25%	0%	0%
	Sur	14%	14%	0%	0%	0%	0%	100%	0%	29%	71%
CSE	Quant	67%	0%	0%	25%	0%	8%	58%	8%	42%	25%
	Qual	20%	0%	0%	0%	20%	40%	40%	20%	40%	40%
	Sur	0%	0%	0%	0%	25%	0%	100%	0%	25%	25%
Totals		38%	11%	3%	11%	3%	9%	47%	7%	24%	30%

identifying high-level data sources. That means that categories in other studies, like attitudinal data described in Randolph et al. [56], was typically grouped in a larger category of "Survey."

The analysis showed that 76% of the papers utilized only one type of data source, even if they collected multiple samples. By only utilizing a single data source, authors limit their understanding of the phenomena under study because they cannot triangulate the results to clarify or understand [16, 41]. The analysis also showed that 30% of all papers used *Surveys* as the sole data source. While surveys are an excellent tool for collecting attitudinal and self-report data, as the only source of data, they may not be robust enough to provide strong evidence to answer a research question. We did not collect data on the use of validated survey instruments as described in Decker and McGill [15].

4.3 RQ3: Norms of Empirical Evaluation

RQ3: Of the papers that have empirical evaluation, do they follow norms (both inclusion and labeling of information needed for replication, meta-analysis, and, eventually, theory-building) for reporting empirical work?

For empirical papers, we expect a paper should contain a research objective, related work, an overview of the research participants, an overview of the study design or methods, a description of the data collection process, a description of the analysis procedures, a report of the results, and a listing of the threats to validity for the research [8, 13, 16, 60] as defined in the *CER Empiricism Assessment − Experimental Rubric*. Any papers reporting the results of a *Survey* as a data source

should have information about how the survey was conducted and the survey questions asked as defined in the *CER Empiricism Assessment – Survey Rubric*. *Descriptive/Persuasive* work should include the goal of the argument, related work, at least three premises, and a conclusion, all with supporting evidence as defined in the *CER Empiricism Assessment – Descriptive/Persuasive Rubric*. However, due to the small numbers of *Descriptive/Persuasive* papers in our analysis (e.g., 10), we do not provide an analysis of that category in this section.

To provide an overall characterization of the papers, we classified each paper into one of three categories for each of the key elements listed above:

- **Strongly Supports Replication** - items that were "Complete and Labeled" provide all the details necessary to support replication;
- **Weakly Supports Replication** - items that were "Complete and Not Labeled" or "Partial and Labeled" provide some, but not all of the details required to support replication;
- **Not Present** - items that were "Not Present" are missing information necessary to support replication.

Table 12 summarizes the results of our analysis of reporting norms. Overall, the only two elements where more than 50% of papers reach the **Strongly Supports Replication** level are *Related Work* and *Results*. Looking across all elements, the papers published in ICER do a better job by having more than 50% of papers reach the **Strongly Supports Replication** level for all elements except for Threats to Validity. However, as the numbers in Table 12 show, there is still room for improvement as a community to reach the **Strongly Supports Replication** level for all attributes most (or all) of the time.

4.3.1 Experimental Rubric. The *Experimental Rubric* describes the standard information we would expect to see in a paper reporting on empirical work. Table 12 summarizes the results for the items on the *Experimental Rubric*.

Research Question. One of the most critical pieces of information in an empirical paper is the research objective, goals, hypotheses, or other summative statement that provides the context of the work reported in the paper. It is concerning that, across all venues, 18% of the papers lacked a summative statement about the work (i.e., scored as **Not Present**). Randolph et al. [56] reported only 22% of papers between 2000 and 2005 had research questions or hypotheses, and Lishinski et al. [36] found that only 47% of CSE and 56% of ICER papers between 2012 and 2015 had explicit research questions. Our numbers for CSE (55%) and ICER (55.3%) for complete and labeled are similar to Lishinski et al. [36] and higher than Randolph et al. [56], suggesting some improvement in the reporting of research questions.

A summative statement guides the reader to the main ideas of the study and is critical for determining if the methods and analysis are appropriate for the statement and if the results answer or address the statement. One or more research statements, objectives, or hypotheses are recommended in standards such as CONSORT [60], APA JARS [7], and AERA [6]. In addition, we found over half of the papers reached only the **Weakly Supports Replication** level because they lacked important information. These papers either did not highlight the summative statement in a meaningful way, through text attributes (e.g., italics or bold), callouts (e.g., boxes or bullets), or clear labeling (e.g., RQ1, Goal), or did not clearly label the goal of the paper, thereby requiring the reader to infer the context of the paper through other statements in the paper.

Related Work. Most papers reported and labeled related work demonstrating **Strongly Supports Replication** for this category. The SIGCSE Board policy[4] states that Program Chairs should

---

Table 12. Norms of Reporting Empirical Evaluation

	Replication Support	SIGCSE TS	ICER	ITiCSE	TOCE	CSE	TOTAL
Research Objective	**Strong**	20.1%	55.3%	17.9%	55.6%	55.0%	28.4%
	Weak	58.0%	31.6%	61.9%	37.0%	45.0%	53.6%
	Not Present	21.9%	13.2%	20.2%	7.4%	0.0%	18.0%
Related Work	**Strong**	62.1%	81.6%	70.2%	81.5%	65.0%	68.0%
	Weak	27.8%	15.8%	19.0%	18.5%	30.0%	23.7%
	Not Present	10.1%	2.6%	10.7%	0.0%	5.0%	8.3%
Subjects	**Strong**	32.0%	50.0%	27.4%	44.4%	44.4%	34.5%
	Weak	60.4%	47.4%	61.9%	51.9%	50.0%	58.0%
	Not Present	7.7%	2.6%	10.7%	3.7%	5.6%	7.4%
Study Design	**Strong**	33.7%	60.5%	56.0%	48.1%	68.8%	45.2%
	Weak	59.2%	36.8%	39.3%	51.9%	31.3%	49.7%
	Not Present	7.1%	2.6%	4.8%	0.0%	0.0%	5.1%
Data Collection	**Strong**	27.8%	52.6%	23.8%	25.9%	60.0%	30.9%
	Weak	65.7%	44.7%	70.2%	66.7%	40.0%	63.4%
	Not Present	6.5%	2.6%	6.0%	7.4%	0.0%	5.7%
Analysis Procedures	**Strong**	28.6%	57.9%	23.8%	48.1%	71.4%	34.1%
	Weak	60.7%	39.5%	65.5%	51.9%	28.6%	57.4%
	Not Present	10.7%	2.6%	10.7%	0.0%	0.0%	8.5%
Results	**Strong**	61.5%	97.4%	78.6%	81.5%	100.0%	73.5%
	Weak	37.3%	2.6%	21.4%	18.5%	0.0%	25.9%
	Not Present	1.2%	0.0%	0.0%	0.0%	0.0%	0.6%
Threats	**Strong**	7.7%	31.6%	4.8%	25.9%	20.0%	11.8%
	Weak	13.0%	23.7%	13.1%	7.4%	20.0%	14.2%
	Not Present	79.3%	44.7%	82.1%	66.7%	60.0%	74.0%

"request that all paper submissions include a review of previous, related work." This guidance demonstrates the importance of establishing reporting norms for the community to ensure high-quality dissemination of empirical results. The papers we rated as **Weakly Supports Replication** typically did not link the related work to the research objective or did not contain an explicitly labeled related work or background section. For those papers that lacked a related work section, the authors typically discussed the related work or background information in the introduction.

Participants. Most papers did include information about the study participants. However, over half of the papers fell into the **Weakly Supports Replication** category because they lacked either full details about the participants or a clear label for the section of the paper describing the participants. A description of participants should contain demographic information to provide context for the results [6, 7, 19, 45, 60]. For example, when papers include student populations, the demographics should include: the number of participants, age groups, education levels, gender, race/ethnicity, region or area of the world, and prior computing experience [45]. Other student demographics that may be of interest, depending on the study, include student disabilities, socioeconomic status, family history (e.g., first-generation college student), and veteran status. Demographics may vary based on the populations under study and the research questions.

In addition to lacking demographics, papers often lacked a description of the recruitment process for the participants. Classroom research may lack the formal control and treatments present in laboratory or controlled studies. By clarifying the inclusion and exclusion criteria for participants, an author clarifies the participant pool and research context [6, 7, 19, 60]. Papers should also include a statement about how the authors recruited participants and obtained consent to participate in the study. This statement demonstrates ethical treatment of research participants and assures readers that authors are following standards for human subject research in their context.

Study Design. Most papers contained a discussion about study design. We found a split between **Strongly Supports Replication** and **Weakly Supports Replication**. There were a number of papers that lacked a clear study design or methods section. In many of these cases, the study design information may have been integrated with the results, rather than as a separate independent section. By providing the study design information in a separate section, the authors help readers more easily extract important information about the design to support replication. In addition to not being labeled, some study designs sections omitted key information. One key piece of information is identification of the dependent and independent variables [7, 60], which could be as simple as identifying the key item under observation or providing overview of an intervention.

Data Collection. Most papers contained a discussion about data collection. However, 63.4% of papers fell into the **Weakly Supports Replication** category. Some of those papers did not include a clearly defined data collection section. Sometimes, the papers integrated the discussion about data collection into the discussion about the study design or methods. Other times, papers integrated the discussion about data collection into the results section. By providing a clear section or subsection around data collection [6, 7], authors can provide details to readers who may be interested in using similar techniques in their own work.

Other **Weakly Supports Replication** papers omitted important information describing the *how*, *where*, and *who* of the data collection. For example, papers frequently lacked details about survey administration, which is discussed further in Section 4.3.2.

Analysis Procedures. Around 34% of papers demonstrated **Strongly Supports Replication** related to the reporting of analysis procedures. Around half of the **Weakly Supports Replication** papers were both incomplete and not properly labeled. These papers were commonly missing a full description of how the authors worked with or processed that data after collection. Additionally, the papers often did not provide a justification for why the particular statistical tests were appropriate for the analysis. A discussion on the use of inferential statistics for ICER's proceedings history provides a justification for full reporting of statistical tests and examples of papers with **Strongly Supports Replication** in this space [58]. Standards support clear reporting on analysis procedures and statistical tests used [6, 7].

Results. Most papers reported results and tied the results back to the research question, goal, or hypothesis (which may have been inferred if not explicitly stated). A large percentage of ICER and TOCE reached the level of **Strongly Supports Replication**. Papers with **Weakly Supports Replication** tended to have partial reporting of results, usually by not tying the results back to the research questions, goals, or hypotheses, as suggested by APA JARS [7]. Some papers also omitted a clearly labeled results section. Standards provide guidance on the details that are included when reporting results [6, 7, 60].

Threats. Most papers (almost 75%) did not report any threats to validity or limitations to work. Our results are similar to the upper range reported by Al-Zubidy et al. [5]. Ihantola et al. [29] found only 22% of papers they reviewed on educational data mining reported threats to validity. If papers did discuss threats or limitations, then the threats/limitations were typically unlabeled and

Table 13. Norms of Reporting Surveys

		SIGCSE TS	ICER	ITiCSE	TOCE	CSE	TOTAL
Conducting the Survey	**Strong Replication**	28%	27%	14%	18%	36%	24%
	Weak Replication	58%	73%	79%	82%	64%	67%
	Not Present	14%	0%	7%	0%	0%	9%
Survey Design	**Strong Replication**	30%	0%	24%	36%	20%	26%
	Weak Replication	50%	91%	50%	45%	20%	51%
	Not Present	20%	9%	26%	18%	60%	23%

appeared in either a results or discussion section. Papers with a **Weakly Supports Replication** may have listed the threats, but typically did not discuss how the authors addressed the threats. A threats to validity section provides context to the study design and provides details on how readers should interpret the work in a broader context and potential biases that might impact the result [5, 7, 60]. Because CER is complex work, threats to validity describe how the authors controlled that complexity (or not) [5, 29]. Replications help control for sampling error and artifacts that identify potential weaknesses with internal validity of the original study. Clearly articulated threats to validity can help identify key variables to vary in replications [59]. Replications also support increasing external validity through generalizability [59].

4.3.2 Survey Rubric. We completed the *Survey Rubric* for any paper that contained a survey as a data source. To provide replicable survey research, authors must provide details on the survey design and survey execution. The reuse and standardization of surveys would be a good starting point for building a culture of replication [6, 41]. If authors provide details about their surveys, even as an un-reviewed supplement to the paper, then it would help build this culture. Table 13 summarizes the results for the items on the Survey Rubric.

Conducting the Survey. We found that most papers demonstrated **Weak Replication**, because they did not fully discuss survey execution in their context. The most frequently missing information was details about survey administration and survey medium. Additionally, many authors did not clearly label the information about conducting the survey.

Survey Design. Most papers did not provide a justification for the selection of survey questions based upon how those questions measure items of interest to answer the research questions, which is a category of reporting in APA JARS [6, 7]. Additionally, most papers did not include the survey questions, which makes it difficult for others to adopt or reuse the survey in their context. In particular, we found that the majority of papers published in CSE during 2014 and 2015 did not discuss the design of the surveys or how the questions were derived, but instead focused more on the distribution and execution of the survey. Validation and adoption of standardized, citable survey instruments will also support replication [41].

5 DISCUSSION

The CER community has seen a growth in submissions and participation. As our community grows, we need to mature the norms of reporting empirical research so the broader community can benefit from the dissemination of high-quality results that answer well-stated research questions, situated in the appropriate literature, with a clear discussion of limitations and threats to validity. Papers should clearly document methods and analysis procedures to allow others to replicate studies in

their own contexts. Reviewers should begin to expect paper authors to follow reporting norms and use those norms as guidance when performing paper reviews.

In the following sections, we provide observations and recommendations from our review of reporting norms. Where appropriate, we connect our recommendations to the broader literature and guidelines, particularly APA JARS [7], CONSORT, [60], *What Works Clearinghouse* [12], AERA [6], and the IES/NSF Guidelines [18, 19].

5.1 Observations and Recommendations from the Results

Based on the results described in the previous section, we make a number of observations and recommendations for authors. We note that these recommendations may not apply in all situations and that the individual research questions and researcher context will impact the decisions that are made during a study.

Evaluation Methods. The results showed that only five papers had used multiple evaluation methods. We encourage researchers to consider using mixed methods approaches more often in study designs to utilize the complementary strengths of both *Qualitative* and *Quantitative* methods (see Fincher and Petre [16], Bishop-Clark and Dietz-Uhler [8], APA JARS [7] for details on mixed method studies). However, we acknowledge that it is not always possible or advisable to use mixed methods approaches and that page limits often constrain the level of detail researchers can include in their publications.

Evaluation Subject Source. The results showed that more than half of the papers used evaluation subjects developed for the specific study. This result suggests that replication in CER is weak. Replication provides a mechanism for generalizing findings about a research question or set of research questions that can lead to the development of theory to support the computing education community [29, 59]. We challenge the computing education community to consider using common methods, evaluation metrics, and data collection and reporting procedures to support the comparison and aggregation of data to generalize CER findings. Utilizing standards can support this goal [6, 7, 12, 19, 60].

Data Source. The results showed that a large number of papers utilized only one source of data. We encourage researchers to utilize multiple data sources to provide more robust insight to the phenomena under study and stronger answers to research questions. In addition, 29% of the papers used surveys as the only data source. There are many situations where surveys are an excellent form of data collection (i.e., understanding a community); however, in classroom studies, researchers should supplement surveys with additional information to provide stronger evidence to answer a research question. For example, a research question about how an intervention impacts student learning should not rely solely on students' self-report of their learning. Margulieux et al. [41] argues that collecting both process (e.g., progress or experience) and product (e.g., performance or outcome) data can increase the applicability of the results to a larger group of educators.

Study Design. We observed a number of papers that lacked a clear study design or methods section. A study design section should tie the observation or intervention to the research question(s) and provide justification that the observation or intervention will support answering the research question(s). The study design section should clearly describe the steps of the study so others could replicate the study in their own environments. For subjective measures, the paper should contain a discussion of the coding process and the method for obtaining agreement or consensus among multiple raters. These items are important for determining how to properly interpret the results. In addition, citing seminal work and standards about the study methods strengthens the methods discussion and can help build the community's knowledge about CER methods [6–8, 12, 16, 60].

Survey Design. We observed that most authors did not provide a justification for the specific survey questions included. While page restrictions can make it difficult to provide the full survey, authors can include the full survey as an un-reviewed supplemental content or on a website. In cases where an author does not want to make the survey publicly available because it may bias future results, the author can include a statement to that effect in the paper and provide contact information for interested researchers to obtain a copy of the survey. The use of standardized survey instruments can support replication [41].

5.2 Author Guidelines for Empirical Papers

Many CER researchers may not have any formal training in conducting human subject research in educational environments. Examples of high-quality literature and study frameworks for replication can help new researchers get started investigating interesting challenges in their own classrooms, departments, and communities. We therefore propose the following guidelines for *CER authors* when preparing their studies and writing up their work. These guidelines are similar to expectations when designing CER studies [8, 16] and have elements in common with standards [6, 7, 12, 19, 60]. CER Venues, like TOCE, are now recommending that authors utilize APA JARS for future submissions.

General Reporting Guidelines

- Clearly identify the evaluation method(s) and subject(s) under evaluation in the study [6, 7, 60].
- Clearly identify what work is new to the study and where the study builds on prior work or data [6, 7, 19].
- Clearly identify any comparisons with prior work or with a control group, if appropriate [6, 7, 19].
- Provide information about participants, including demographics [6, 7, 45, 60].
- Identify the study as observational or interventional [7].
- Identify data sources [6, 7].
- For surveys, describe the administration process and provide the survey questions.
- Where possible, provide supplemental resources that would help others replicate the work, e.g., code used to analyze data on a public version control system or anonymized, aggregate data on a website [19, 60].
- Utilize pre-registration and open science to strengthen research integrity and transparency [19, 42, 49, 60].

Research Question [6, 7, 60]

- The research goal or question is of upmost importance because it drives the rest of the work. Utilize resources on writing high-quality and actionable research questions.
- Highlight the research goal and questions in the text. For example, italicize the goal statements in the paper's introduction. List and name more specific research questions (e.g., RQ1, RQ2) in the introduction and revisit them in the results.

Related Work

- Report related work in its own section. It should situate the research question or goal in the context of the broader literature. Because *most* SIGCSE conferences now provide extra pages for references, authors should be able to provide a robust discussion of related work with sufficient citations of relevant literature.

- Provide a discussion that links the related work to the research goal, objectives, or questions for the paper. The section should also discuss how the related work informs the study and how the study builds on previous work. The information about previous work is especially important for replication papers [6, 7].
- Synthesize the key themes in the prior work to provide context and motivation for the current study, which should not simply be an annotated bibliography.

Participants

- Report demographic information about the participant groups in the study. For students, minimally report numbers, ages, education levels, gender, race/ethnicity, prior computing experience, and regional location for the students [6, 7, 45]. McGill et al. [45] provide other suggestions for reporting demographic information for other participant populations.
- If number of participants is too small that reporting demographics would become identifying for those participants, then clearly state this information in lieu of reporting specific demographics.
- Describe the process for selecting participants. If the study uses multiple groups, then the paper should discuss the process for allocating participants to each group. If the authors exclude participants from the study (e.g., minors in a university-level study), then the paper should explain the exclusion criteria [6, 7, 60].
- Explain the process for obtaining consent from participants and any ethical considerations associated with human subjects research. Since ethical standards vary by country and institution, authors should clarify expectations in their context and assure reviewers and readers that the human subjects study was handled in an appropriate manner for the author's context [6, 7].

Study Design

- Separate the study design or methods from the discussion of results. This separation will allow readers to assess the quality of the study and to more easily see the steps needed for replication.
- Identify independent and dependent variables [7, 60].
- Describe each step in conducting the study [6, 7, 60].

Data Collection

- Identify data collection procedures, including mediums for data collection, who collected the data, and how the data addresses or answers the research questions [6, 7, 19].
- Provide a rationale for why the data collected will help address the research question(s) [6, 7].

Analysis Procedures

- Describe the process for working with the collected data, including the process for cleaning data, if necessary [6, 19].
- For qualitative analysis, describe the coding process and the process for evaluating the correctness of the coding (e.g., multiple raters, inter-rater reliability) [6, 7].
- For quantitative analysis, describe, and justify, the statistical tests, if appropriate. For other types of analysis, provide a justification on how the analysis helps answer the research question(s) [6, 7, 60].

Results

- Ensure that results directly address the research question(s) [6, 7, 60].
- For qualitative results, provide tables, descriptions, quotes, and arguments to answer the research questions [6, 7].
- For quantitative results, provide summary or descriptive statistics to answer the research questions [6, 7].

Threats to Validity

- Include a dedicated threats to validity section that lists internal, external, construct threats, and biases as appropriate for the research question(s) and study design [6, 7, 60].
- Discuss how the study design addresses the threats.
- Discuss any threats not addressed by the study design.
- Explain and justify the unaddressed threats.
- Explain the potential impact on the interpretation of the study results due to the unaddressed threats [6, 7, 60].

5.3 Reviewer Guidelines for Reviewing Empirical Papers

The responsibility for high-quality publications about CER work does not rest solely on authors. Review guidelines over the past several years have clarified expectations for reviewing empirical work, both as CER and experience reports. Reviewers have a responsibility to hold authors accountable for following norms for reporting CER. We recommend that reviewers use the author guidelines and standards documents above as a checklist for things to provide feedback on during peer review. These guidelines and standards echo guidelines and standards from education and other fields about presentation of empirical work [6, 7, 12, 19, 60].

6 THREATS TO VALIDITY

In developing our rubric and applying it to the work described in this article, we identified some limitations and threats to validity we must address. We have backgrounds in conducting human-based empirical research in computing education, software engineering, and security, but do not explicitly have research degrees in education. However, two of authors have been members of the program committee of SIGCSE TS for multiple years, including years volunteering as program and symposium chairs, so we believe that we do have the needed background.

During the creation of the *CER Empiricism Assessment Rubric*, we did not explicitly review or reference other validated rubrics. Therefore, it is possible that we omitted some aspects of educational research project design. To address this threat to validity, after we completed our work, we compared our rubric to other guidelines for reporting and assessing empirical research study quality, such as APA JARS [7], the *What Works Clearinghouse Standards Handbook* [12], and CONSORT [60], as discussed in Section 2.2. We found that there was significant overlap in the core concepts of these other guidelines and rubrics with our own, indicating that our rubric captures many of the same reporting values.

In choosing the years and venues for inclusion in this review, we selected the conferences and journals that are widely considered to be the top tier for CER and only analyzed the proceedings or issues from two years, 2014 and 2015. We did not include more general education conferences that have a computing track, such as ASEE or FIE, and we excluded some other venues (e.g., Koli Calling) to make the review more feasible. While this choice means we could have missed a portion of the community, we believe that the venues we chose cover the majority of the computing education literature. Further, some other venues accepted papers based solely on the abstract, rather than the

entire paper, which we believe does not accurately represent the current state of empiricism. We chose the particular set of years to create a baseline on empirical work before more recent changes to paper tracks and reviewing guidelines at SIGCSE TS and other venues. Also, we recognize that the findings from 2014 and 2015 may not represent the current state of empiricism in 2021. Future work will consider a review of more recent publications.

Furthermore, we do not have the perspective of manuscripts that were not accepted into these venues. As Fincher et al. discussed in the conclusion of their book chapter [17], a weakness of systematic literature reviews is that they focus on a quantitative analysis to characterize a body of work through an author-defined lens and may miss the broader context of the timeframe in which the work was published. We do not consider the broader context of computing education during 2014 and 2015 and how that may impact the types of papers accepted. However, we minimize this limitation by focusing on characterizing empirical elements independent of the specific topic of the work (e.g., K–12 or CS1). Our work provides value by suggesting how the community can improve reporting standards independent of the broader community context for the acceptance and publication of the reviewed literature.

In our application of the rubric, we settled on a two-pass approach. Two researchers evaluated each paper using the *Base Rubric*—evaluation method, evaluation subject, evaluation subject source, comparison, and number of participants. We used this approach to better ensure proper overall categorization of papers. We calculated the inter-rater reliability among the research team working on the categorization and determined only minor differences in the evaluations. After this initial evaluation, one researcher then completed the remaining items in the *CER Empiricism Assessment Rubric* sub-rubrics as appropriate for each paper. We chose this approach because it gave us the best balance between efficiency and accuracy.

However, we recognize the possibility that individual researchers mischaracterized aspects of individual papers. For example, one researcher might rate *General Rubric: Research Objectives* for a paper as *Partial and Not Labeled* because she found a discussion of research objectives, but not a dedicated section, while another reader could have missed the discussion altogether because he read too quickly. We have provided our dataset [27] for other researchers to consider when utilizing our rubric.

7 CONCLUSION

Our research goal was to characterize the reporting of empiricism in Computing Education Research literature by identifying whether publications include content necessary for researchers to perform replications, meta-analyses, and theory building. This systematic literature review summarizes the type of papers and studies included during 2014 and 2015 in the SIGCSE TS, ICER, ITiCSE, TOCE, and CSE venues. A majority of the accepted papers report empirical work. However those papers do not consistently follow reporting norms. We have provided suggestions to authors and reviewers to move the community forward in publishing high-quality empirical work that can lead to meta-analysis and theory building.

We did observe progress in reporting empirical work in recent years. With the creation of the CS education research track at the SIGCSE TS for SIGCSE 2018 [54], the organizers updated the review criteria to specifically request reviewers evaluate the items we include in our *General Rubric*. Additionally, TOCE now recommends that authors utilize APA JARS when organizing their submissions.

In future work, we will conduct a similar literature review on more recent publications to gauge whether the community has made any progress. With improved reporting, our next review will attempt to consider categorizing papers at a more granular level than survey, qualitative,

and quantitative, which would provide the opportunity for more detailed rubrics. We welcome feedback on our rubric for possible revisions in this future review.

As computing education is growing as a field and community, we need to establish norms for reporting empirical work. By doing so, we will support replication and meta-analysis. Increased rigor in reporting expectations will increase the reputation of CER in the broader computing research community, which will facilitate the growth and increased reputation of CER scholars and the computing education field. We all need to contribute: authors and researchers need to create and report well-designed, high-quality research studies or well-documented and supported experience reports; reviewers need to provide feedback not only on the novelty of the idea, but the quality of presentation; and the community needs to support replications and meta-analyses so we can grow our understanding of how to share computing education with the world.

APPENDIX A: CER EMPIRICISM ASSESSMENT RUBRIC

INTRODUCTION

The CER Empiricism Assessment Rubric was developed as a repeatable methodology for determining the degree and rigor in which empirical research principles are reported regarding research activity in CER literature. The rubric makes no judgment on the quality of the research or of the research question of the reported project. Instead, the rubric tries to establish the effectiveness of the paper at communicating the various aspects of the research activity to the reader such that it could not only be understood, but possibly replicated for further investigation.

RUBRIC QUESTION RESPONSE VALUES

For all sub-rubrics other than the Base Rubric, we only use values from the following scale, as discussed in Section 3.2:

- Complete and Labeled
- Complete and Not labeled
- Partial and Labeled
- Partial and Not Labeled
- Not Present
- Not Applicable

The descriptive values address two key two parts of reporting CER work—the completeness of the presentation of the information and whether the information is labeled in the paper or not.

Completeness - the level of completeness of presented information

- Complete - Answers/addresses all questions for a rubric category. There is no assessment on the quality of the answer.
- Partial - Answers/addresses some of the questions for a rubric category. There is no assessment on the quality of the answer.
- Not Present - Answers/addresses none of the questions for a rubric category. The questions should be addressed.
- Not Applicable - The rubric item is not applicable to the paper.

Labeled - whether the presented information is clearly labeled

- Labeled - There is a heading appropriate for the rubric item or there is emphasis (bold/italics) for the rubric item.

• Not Labeled - There is no heading or emphasis for the rubric item to easily find the item in the paper.

RUBRIC STEPS

Follow the steps highlighted in the boxes. Key sub-activities associated with each step in the rubric are denoted by §.

> **Step 0: Read the research work you wish to evaluate.**

> **Step 1: Apply the Base Rubric to the research work.**

(BR) BASE RUBRIC

The Base Rubric consists of five high-level questions that provide basic information about a work, including the primary methodology, the subject being studied, where the subject originated, whether there is any comparison, and the size of the study as measured by the number of participants or data instances.

(BR-1) Base Rubric: Evaluation Method

After reading the paper, the first step is to determine the evaluation method used in the work. The evaluation method is a high-level categorization regarding the overall nature of the project that is being reported. It is possible to select multiple options for the evaluation method, although any more than two would be highly unusual. An example of this could be a project that examined both student assessment scores and student survey responses when evaluating a new teaching method. In this case, the project could be categorized as *Experimental, Survey*.

Literature Review. A work is categorized as a literature review if it is mainly focused on reporting on the current state of the body of knowledge on a particular topic or research question. The work does not necessarily have to add to the body of knowledge, but it often will draw conclusions on where the state of the research is heading.

§: If the work is a literature review, complete item BR-2 (Evaluation Subject) from the Base Rubric (BR) only and then continue with Descriptive/Persuasive Rubric (DR).

Exploratory. An exploratory work is a research work-in-progress. An exploratory project could originate from a model building exercise, observation without any predefined research questions, or building a framework or taxonomy.

§: If the work is exploratory, complete the Base Rubric (BR) and the Experimental Rubric (ER).

Descriptive/Persuasive. A descriptive or persuasive work describes a current situation or paints a picture but does not test predictions nor does it imply any cause-and-effect relationships. This type of work is different than a literature review in that it is reporting on a research subject rather than the current body of knowledge.

§: If the work is descriptive/persuasive, complete the Base Rubric (BR) and the Descriptive/Persuasive Rubric (DR).

Survey. Survey works use information gathered through forms or some other asynchronous querying method as their primary data source. CER project often have surveys or similar instruments (such as student evaluations) as a secondary data source even when there are other data sources being used.

§: If the work uses a survey as one of its evaluative methods, complete the Base Rubric (BR), the Experimental Rubric (ER), and the Survey Data Source Rubric (SR).

Qualitative. The primary indicator of an evaluative project with a qualitative evaluation method is the presence of free form or free text answers from participants. These responses then have to be individually coded or evaluated separately to draw any conclusions regarding the research questions of the project. Qualitative data is common for projects with a small sample size, where in-person interviews or focus groups are used, or if participants are audio or video recorded in any way. In each of these cases, the data typically requires more time to evaluate, code, or process as opposed to quantitative data.

§: If the work uses qualitative data as one of its evaluative methods, complete the Base Rubric (BR) and the Experimental Rubric (ER).

Quantitative. Whereas qualitative data is defined generally by free form information gathered from participants, quantitative data is identified by discrete values and counts that can be more easily and directly compared with each other. While it can be easier to identify a quantitative study that has defined control and treatment groups with a large amount of data, case studies, experience reports, and quasi-experimental studies can also be considered quantitative studies, depending on the types of data collected. Quantitative data in CER is often gathered from Likert-type questions on participant surveys, course assessment data, enrollment and retention data, demographic data, and other information that could inform some quality or aspect of a pedagogical technique, course, or curricula.

§: If the work uses quantitative data as one of its evaluative methods, complete the Base Rubric (BR) and the Experimental Rubric (ER).

Missing. A work that states a causal relationship as fact but does not support that claim with any empirical research or data is considered to be missing the evaluation method.

§: If the work's evaluation method is missing, there is no further categorization to be completed.

Not Applicable. Some CER venues publish works that require no evaluation, nor are taking any particular stance. These works could, for example, simply describe a new course or a new curriculum with no claims of effectiveness.

§: If the work does not require any empirical evaluation, there is no further categorization to be completed.

Mixed Methods. Any work that explicitly is using more than one method listed above is considered a mixed methods study.

§: If the work is considered a mixed methods study, denote each method in a comma-separated list and then complete all appropriate rubrics as listed with the method.

(BR-2) Base Rubric: Evaluation Subject

The evaluation subject rubric item identifies the nature of the treatment that is being investigated by the work. In a well-focused research work, there should only be one evaluation subject.

§: Select one of the following options as a part of the Base Rubric (BR).

Pedagogical Technique. A work is evaluating a pedagogical technique if the researchers are particularly interested in a specific teaching method. Studies that are evaluating a pedagogical technique often focus on student learning outcomes as an indicator of the effectiveness of the treatment, which could be represented as assessment information or student perceptions of their learning in a course or for a particular knowledge unit. These studies are also more likely to have threats to validity from instructor bias and should be addressed appropriately.

Tool. A work is evaluating a tool if the research questions are addressing the effectiveness of the tool itself and not the underlying pedagogical technique. This could also encompass various forms of educational technology, such as the effectiveness of distance learning tools. While there is almost undoubtedly a pedagogical approach associated with a tool, selecting this option as the evaluation subject indicates that the implementation and usage of the tool itself is the primary focus.

Curriculum. A curriculum work is a research project that is looking beyond just one particular course, but rather the creation and integration of multiple units or courses across a larger program. This could also include special curricula, such as summer camps.

Assessment. A research project where an assessment is the evaluation subject is focused on a particular assignment or quiz or set of assignments within a single knowledge unit or course. These types of projects often look at types of assessments and what makes them effective, validating assessments, issues with scale, and academic integrity.

Community. Community-based research projects attempt to ascertain some understanding of a group of individuals. For example, a project that is trying to determine the relative preparedness of underrepresented minority students before attending college could be categorized as community.

Other or Combination. Other research works could fall outside the range of this rubric, or could be considered a combination of the items above.

(BR-3) Base Rubric: Evaluation Subject Source

The evaluation subject source rubric item establishes where the evaluation subject was first created. In our examination of CER literature, we noticed that replication studies did not occur often or it was sometimes unclear if the treatment used was being introduced in the current paper of if this was a continuation of previous work. Thus, this rubric item denotes whether the evaluation subject was created originally by the authors or elsewhere and whether the treatment was modified for use in the current study.

§: Select one of the following options as a part of the Base Rubric (BR).

Authors Here. An evaluation subject is considered to be in this category if the authors created the treatment themselves for use in this study.

Authors Elsewhere. If a subject was created by the authors elsewhere, then the treatment was first presented in an previous published work.

Other Modified. A subject that was created by someone other than the authors and has been altered in some way would fit in this category.

Other Not Modified. If the current work is a true replication study using a treatment that has not been altered and was created by another person, then this category is selected.

Community. If the Evaluation Subject has been identified as *Community*, then the Evaluation Subject Source will also be *Community*, as it represents the idea that the subject is a self-identifying group of people related to an area of interest.

(BR-4) Base Rubric: Comparison

Sometimes a research question is intended to discover and report on the current state of the world. However, many research questions aim to determine whether a treatment had an effect on

a population. If this is the case, then the work should compare the results in some way to some form of control data.

§: *Select one of the following options as a part of the Base Rubric (BR).*

Historical. If the results are compared to data from before the treatment, then indicate that the comparison is historical in nature.

Comparison. If the results are compared to data generated as a part of the current research study, such as a from a specific control group or A/B testing, then select this category.

None. If the results are reporting on the state of the world and are not compared to any other dataset, then indicate that there was no comparison.

(BR-5) Base Rubric: Number of Participants

The number of participants indicates the *n* value of the study. This could be the number of students in a course, the number of submissions to a grading system, the number of responses to a survey, and so on.

§: *Select one of the numeric range options as a part of the Base Rubric (BR).*

- None
- 1
- 2
- 3–10
- 11–30
- 31–75
- 76–150
- 151–300
- 301–999
- 1,000+

> **Step 2: If the work was categorized as *Exploratory, Survey, Qualitative*, or *Quantitative*, complete the Experimental Rubric (ER).**

(ER) EXPERIMENTAL RUBRIC

For each of these rubric items, answer using the guidelines listed in the Rubric Question Response Values section unless otherwise directed. Each item has one or more questions that can aid you in determining the correct rubric value for the overall item.

§: *For each Experimental Rubric (ER) item, select one value from the Rubric Question Response Values that best describes how that particular item was reported in the paper unless otherwise directed. Use the sub-questions with each item to help identify the proper selection.*

(ER-1) Experimental Rubric: Research Objectives

- Does the paper include a description of the research objectives? (e.g., goals, questions, hypotheses)

(ER-2) Experimental Rubric: Related Work

- Does the paper present related work?
- Does the paper link the research objectives directly to the related work?

(ER-3) Experimental Rubric: Participants

- Does the paper provide demographics on the participants?
- Does the paper describe the sampling/recruitment methods used? (i.e., why these respondents? e.g., mailing list, advertised, etc.)

(ER-4) Experimental Rubric: Study Design

- Does the paper define the dependent and independent variables (including specific metrics)?
- Does the paper justify the variables in relevance to the overall research objective?
- Does the paper describe the treatments/protocol/steps followed in the study?
- For subjective measure, does the paper describe any type of inter-rater agreement?

(ER-5) Experimental Rubric: Data Collection

- Does the paper describe who gathered the data?
- Does the paper describe how the data was gathered?
- Does the paper describe where the data was gathered?

(ER-6) Experimental Rubric: Analysis Procedures

- Does the paper describe the analysis procedures? - process for working with the data after collection
- For Qualitative Data:
  — Does the paper describe how they cleaned/coded the data?
  — Does the paper describe how they evaluated the correctness of the coding (e.g., inter-rater reliability)?
- For Quantitative Data:
  — Does the paper describe the specific statistical tests that are used to analyze the data? (e.g., hypothesis checks, statistical tests, p-values, performance metrics, precision, recall, accuracy, False positive, False negative, etc.)
  — Does it justify why the tests were chosen?

(ER-7) Experimental Rubric: Results

- Does the paper include summary/descriptive statistics? (e.g., mean, std dev, charts/tables to describe data)
- Does the paper discuss results in relation to the research objectives? (e.g., hypotheses evaluated, questions answered, or "big picture")
- Does the paper present the qualitative data, if applicable? (tables, descriptions, arguments)

(ER-8) Experimental Rubric: Threats to Validity

- Does the paper contain a dedicated discussion of the threats to validity (i.e., limitations or mitigations)?
- Does this section include a discussion of how the threats were addressed?
- Does this section include a discussion of threats left unaddressed?

(ER-9) Experimental Rubric: Type of Study

For this rubric item, determine whether the treatment is classified as observational or interventional in nature.

§: Select one of the following options:

- Observational - Study is performed in a natural setting in which the researcher collects data via observation without manipulation of the situation. In this type of study, the researcher is merely observing the participants in a natural setting without interacting with the participants.
- Interventional - Study is performed by assigning participants into groups (e.g., control and experimental) and a treatment is applied to the experimental group to measure its effect. In this type of study, the researcher is interacting with the participants to study their response to certain variables.

(ER-10) Experimental Rubric: Data Sources

§: Select all sources of data used in the study from this list.

- Survey
- Course Evaluations
- Assessment Data
- Observations
- Interviews
- Focus Groups
- Automated
- Logs/Metadata/Generated/Mined
- Other

> **Step 3: If the work was categorized as *Survey*, complete the Survey Data Source Rubric (ER).**

(SR) SURVEY DATA SOURCE RUBRIC

Complete this section if the evaluation method for the work was classified as survey or if one of the selected data sources for an experimental study was a survey. The purpose of this rubric is to determine the quality of how information regarding the survey instrument was developed and presented.

§: For each Survey Data Source Rubric (SR) item, select one value from the Rubric Question Response Values that best describes how that particular item was reported in the paper. Use the sub-questions with each item to help identify the proper selection.

(SR-1) Survey Data Source Rubric: Conducting the Survey

- Does the paper describe the study design? (e.g., pre- and post-surveys, reflections after assignments, etc.)
- Does the paper describe how the survey was administered? (e.g., in-person, remote)
- Does the paper describe the survey medium?

(SR-2) Survey Data Source Rubric: Survey Design

- Does the paper provide a rationale behind the questions? (i.e., why these questions and not others)
- Does the paper include the survey questions or provide a link to them?

> **Step 4: If the work was categorized as *Descriptive/Persuasive*, complete the Descriptive/Persuasive Rubric (DR).**

(DR) DESCRIPTIVE/PERSUASIVE RUBRIC

Some papers that are published are not full research studies, but rather describe a current situation and do not imply any cause-and-effect relationships. If this is the case, complete the following rubric items using the Rubric Question Response Values.

§: *For each Descriptive/Persuasive Rubric (DR) item, select one value from the Rubric Question Response Values that best describes how that particular item was reported in the paper. Use the subquestions with each item to help identify the proper selection.*

(DR-1) Descriptive/Persuasive Rubric: Goal of the Argument

- Does the paper describe the goal of the argument?

(DR-2) Descriptive/Persuasive Rubric: Related Work (Context)

- Does the paper present related work?
- Does the paper link the research objectives directly to the argument?

(DR-3) Descriptive/Persuasive Rubric: Premises and a Conclusion

- Does the paper contain two or more premises and a conclusion? (Aristotle's rule)

(DR-4) Descriptive/Persuasive Rubric: Supporting Evidence

- Does the paper describe the supporting evidence?
- Does the paper provide references for the the supporting evidence?

REFERENCES

[1] Quintin Cutts, Beth Simon, and Brian Dorn. 2014. *Proceedings of the 10th Annual Conference on International Computing Education Research.* Association for Computing Machinery, New York, NY.

[2] 2015. *Proceedings of the 11th Annual International Conference on International Computing Education Research.* Association for Computing Machinery, New York, NY.

[3] ACM. 2020. Artifact Review and Badging – Current. Retrieved from https://www.acm.org/publications/policies/artifact-review-and-badging-current.

[4] Alireza Ahadi, Arto Hellas, Petri Ihantola, Ari Korhonen, and Andrew Petersen. 2016. Replication in computing education research: Researcher attitudes and experiences. In *Proceedings of the 16th Koli Calling International Conference on Computing Education Research (Koli Calling'16).* ACM, New York, NY, 2–11. DOI: https://doi.org/10.1145/2999541.2999554

[5] Ahmed Al-Zubidy, Jeffrey C. Carver, Sarah Heckman, and Mark Sherriff. 2016. A (updated) review of empiricism at the SIGCSE Technical Symposium. In *Proceedings of the 47th ACM Technical Symposium on Computing Science Education (SIGCSE'16).* Association for Computing Machinery, New York, NY, 120–125. DOI: https://doi.org/10.1145/2839509.2844601

[6] American Educational Research Association. 2006. Standard for reporting on empirical social science research in AERA publications. *Educ. Res.* 35, 6 (2006), 33–40. DOI: https://doi.org/10.3102/0013189X035006033

[7] American Psychological Association. 2018. Journal Reporting Standards (JARS). Retrieved from https://apastyle.apa.org/jars.

[8] Cathy Bishop-Clark and Beth Dietz-Uhler. 2012. *Engaging in the Scholarship of Teaching and Learning.* Stylus Publishing, Sterling, VA.

[9] Paulo Blikstein and Sepi Hejazi Moghadam. 2019. *Computing Education Literature Review and Voices from the Field.* Cambridge University Press, 56–78. DOI : https://doi.org/10.1017/9781108654555.004

[10] Jeffrey C. Carver. 2010. Towards reporting guidelines for experimental replications: A proposal. In *Proceedings of the 1st International Workshop on Replication in Empirical Software Engineering*, Vol. 1. Citeseer, 1–4.

[11] Tony Clear. 2006. Valuing computer science education research? In *Proceedings of the 6th Baltic Sea Conference on Computing Education Research: Koli Calling 2006 (Baltic Sea'06).* ACM, New York, NY, 8–18. DOI : https://doi.org/10.1145/1315803.1315806

[12] What Works Clearinghouse. 2020. What Works Clearinghouse Standards Handbook, Version 4.1. Retrieved from https://ies.ed.gov/ncee/wwc/Handbooks.

[13] John W. Creswell and J. David Creswell. 2018. *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches* (5th ed.). SAGE Publications, Inc.

[14] Mats Daniels and Arnold Pears. 2012. Models and methods for computing education research. In *Proceedings of the 14th Australasian Computing Education Conference (ACE'12).* Australian Computer Society, Inc., AUS, 95–102.

[15] Adrienne Decker and Monica M. McGill. 2019. A topical review of evaluation instruments for computing education. In *Proceedings of the 50th ACM Technical Symposium on Computer Science Education (SIGCSE'19).* Association for Computing Machinery, New York, NY, 558–564. DOI : https://doi.org/10.1145/3287324.3287393

[16] Sally Fincher and Marian Petre (Eds.). 2004. *Computer Science Education Research.* Taylor & Francis.

[17] Sally A. Fincher, Josh Tenenberg, Brian Dorn, Christopher Hundhausen, Robert McCartney, and Laurie Murphy. 2019. *Computing Education Research Today.* Cambridge University Press, 40–55. DOI : https://doi.org/10.1017/9781108654555.003

[18] Institute for Education Science and the National Science Foundation. 2013. Common Guidelines for Education Research and Development. Retrieved from https://www.nsf.gov/pubs/2013/nsf13126/nsf13126.pdf.

[19] The National Science Foundation and the Institute of Education Sciences. 2018. Companion Guidelines on Replication & Reproducibility in Education Research: A Supplement to the Common Guidelines for Education Research and Development. Retrieved from https://www.nsf.gov/pubs/2019/nsf19022/nsf19022.pdf.

[20] Varvara Garneli, Michail N. Giannakos, and Konstantinos Chorianopoulos. 2015. Computing education in K–12 schools: A review of the literature. In *Proceedings of the IEEE Global Engineering Education Conference (EDUCON'15).* IEEE, 543–551. DOI : https://doi.org/10.1109/EDUCON.2015.7096023

[21] Omar S. Gómez, Natalia Juristo, and Sira Vegas. 2010. Replication, reproduction and re-analysis: Three ways for verifying experimental findings. In *Proceedings of the 1st International Workshop on Replication in Empirical Software Engineering Research (RESER'10).*

[22] Mark Guzdial and Benedict du Boulay. 2019. *The History of Computing Education Research.* Cambridge University Press, 11–39. DOI : https://doi.org/10.1017/9781108654555.002

[23] A. B. Haidich. 2010. Meta-analysis in medical research. *Hippokratia* 14 (Suppl 1) (2010), 29–37.

[24] Qiang Hao, David H. Smith IV, Naitra Iriumi, Michail Tsikerdekis, and Amy J. Ko. 2019. A systematic investigation of replications in computing education research. *ACM Trans. Comput. Educ.* 19, 4 (Aug. 2019). DOI : https://doi.org/10.1145/3345328

[25] Elizabeth K. Hawthorne, Manuel A. Pérez-Quiñones, Sarah Heckman, and Jian Zhang. 2019. SIGCSE Technical Symposium 2019 report. *SIGCSE Bull.* 51, 2 (Apr. 2019), 2–4. DOI : https://doi.org/10.1145/3329103.3329104

[26] Orit Hazzan, Yael Dubinsky, Larisa Eidelman, Victoria Sakhnini, and Mariana Teif. 2006. Qualitative research in computer science education. *ACM SIGCSE Bull.* 38, 1 (2006), 408–412.

[27] Sarah Heckman, Jeffrey C. Carver, Mark Sherriff, and Ahmed Al-Zubidy. 2021. CER Empiricism Assessment Dataset (2014 & 2015). (June 2021). DOI : https://doi.org/10.6084/m9.figshare.14816634.v1

[28] Christian Holmboe, Linda McIver, and Carlisle E. George. 2001. Research agenda for computer science education. In *Proceedings of the 13th Psychology of Programming Interest Group (PPIG) Workshop.* Retrieved from https://www.ppig.org/files/2001-PPIG-13th-holmboe.pdf.

[29] Petri Ihantola, Arto Vihavainen, Alireza Ahadi, Matthew Butler, Jürgen Börstler, Stephen H. Edwards, Essi Isohanni, Ari Korhonen, Andrew Petersen, Kelly Rivers, et al. 2015. Educational data mining and learning analytics in programming: Literature review and case studies. In *Proceedings of the ITiCSE Working Group Reports.* 41–63. DOI : https://doi.org/10.1145/2858796.2858798

[30] John Impagliazzo, Ming Zhang, and Xi Wu. 2018. SIGCSE launches new conference on a global scale. *SIGCSE Bull.* 50, 4 (Oct. 2018), 2–3. DOI : https://doi.org/10.1145/3287087.3287088

[31] Andreas Jedlitschka and Dietmar Pfahl. 2005. Reporting guidelines for controlled experiments in software engineering. In *Proceedings of the International Symposium on Empirical Software Engineering.* IEEE. DOI : https://doi.org/10.1109/ISESE.2005.1541818

[32] Mike Joy, Jane Sinclair, Shanghua Sun, Jirarat Sitthiworachart, and Javier López-González. 2009. Categorising computer science education research. *Educ. Inf. Technol.* 14, 2 (2009), 105–126. DOI : https://doi.org/10.1007/s10639-008-9078-4

[33] Päivi Kinnunen, Veijo Meisalo, and Lauri Malmi. 2010. Have we missed something?: Identifying missing types of research in computing education. In *Proceedings of the 6th International Workshop on Computing Education Research*. ACM, 13–22. DOI : https://doi.org/10.1145/1839594.1839598

[34] Barbara Kitchenham. 2004. *Procedures for Performing Systematic Reviews*. Joint Technical Report, Keele University Technical Report TR/SE-0401, NICTA Technical Report 040011T.1. Retrieved from https://www.inf.ufsc.br/~aldo.vw/kitchenham.pdf.

[35] Barbara Kitchenham, Hiyam Al-Khilidar, Muhammed Ali Babar, Mike Berry, Karl Cox, Jacky Keung, Felicia Kurniawati, Mark Staples, He Zhang, and Liming Zhu. 2008. Evaluating guidelines for reporting empirical software engineering studies. *Empir. Softw. Eng.* 13, 1 (2008), 97–121. DOI : https://doi.org/10.1007/s10664-007-9053-5

[36] Alex Lishinski, Jon Good, Phil Sands, and Aman Yadav. 2016. Methodological rigor and theoretical foundations of CS education research. In *Proceedings of the ACM Conference on International Computing Education Research*. 161–169. DOI : https://doi.org/10.1145/2960310.2960328

[37] Andrew Luxton-Reilly, Ibrahim Albluwi, Brett A. Becker, Michail Giannakos, Amruth N. Kumar, Linda Ott, James Paterson, Michael James Scott, Judy Sheard, and Claudia Szabo. 2018. Introductory programming: A systematic literature review. In *Proceedings of the 23rd Annual ACM Conference on Innovation and Technology in Computer Science Education*. 55–106. DOI : https://doi.org/10.1145/3293881.3295779

[38] Lauri Malmi, Judy Sheard, Päivi Kinnunen, Simon, and Jane Sinclair. 2020. Theories and models of emotions, attitudes, and self-efficacy in the context of programming education. In *Proceedings of the ACM Conference on International Computing Education Research*. 36–47. DOI : https://doi.org/10.1145/3372782.3406279

[39] Lauri Malmi, Judy Sheard, Simon, Roman Bednarik, Juha Helminen, Päivi Kinnunen, Ari Korhonen, Niko Myller, Juha Sorva, and Ahmad Taherkhani. 2014. Theoretical underpinnings of computing education research: What is the evidence? In *Proceedings of the 10th Annual Conference on International Computing Education Research*. 27–34. DOI : https://doi.org/10.1145/2632320.2632358

[40] Lauri Malmi, Simon, Judy Sheard, Roman Bednarik, Juha Helminen, Ari Korhonen, Niko Myller, Juha Sorva, Ahmad Taherkhani, et al. 2010. Characterizing research in computing education: A preliminary analysis of the literature. In *Proceedings of the 6th International Workshop on Computing Education Research*. ACM, 3–12. DOI : https://doi.org/10.1145/1839594.1839597

[41] Lauren Margulieux, Tuba Ayer Ketenci, and Adrienne Decker. 2019. Review of measurements used in computing education research and suggestions for increasing standardization. *Comput. Sci. Educ.* 29, 1 (2019), 49–78. DOI : https://doi.org/10.1080/08993408.2018.1562145

[42] Monica M. McGill. 2019. Discovering empirically-based best practices in computing education through replication, reproducibility, and meta-analysis studies. In *Proceedings of the 19th Koli Calling International Conference on Computing Education Research (Koli Calling'19)*. Association for Computing Machinery, New York, NY. DOI : https://doi.org/10.1145/3364510.3364528

[43] Monica M. McGill and Adrienne Decker. 2018. Defining requirements for a repository to meet the needs of K–12 computer science educators, researchers, and evaluators. In *Proceedings of the IEEE Frontiers in Education Conference (FIE'18)*. 1–9. DOI : https://doi.org/10.1109/FIE.2018.8659312

[44] Monica M. McGill and Adrienne Decker. 2020. Construction of a taxonomy for tools, languages, and environments across computing education. In *Proceedings of the ACM Conference on International Computing Education Research*. 124–135. DOI : https://doi.org/10.1145/3372782.3406258

[45] 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*. ACM, 964–969. DOI : https://doi.org/10.1145/3159450.3159481

[46] Rodrigo Pessoa Medeiros, Geber Lisboa Ramalho, and Taciana Pontual Falcão. 2018. A systematic literature review on teaching and learning introductory programming in higher education. *IEEE Trans. Educ.* 62, 2 (2018), 77–90. DOI : https://doi.org/10.1109/TE.2018.2864133

[47] Diba Mirza, Phillip T. Conrad, Christian Lloyd, Ziad Matni, and Arthur Gatin. 2019. Undergraduate teaching assistants in computer science: A systematic literature review. In *Proceedings of the ACM Conference on International Computing Education Research*. 31–40. DOI : https://doi.org/10.1145/3291279.3339422

[48] Greg L. Nelson and Amy J. Ko. 2018. On use of theory in computing education research. In *Proceedings of the ACM Conference on International Computing Education Research*. 31–39. DOI : https://doi.org/10.1145/3230977.3230992

[49] Brian A. Nosek and Daniël Lakens. 2014. Registered reports: A method to increase the credibility of published results. *Soc. Psychol.* 45, 3 (2014), 137–141. DOI : https://doi.org/10.1027/1864-9335/a000192

[50] National Academies of Sciences, Engineering, and Medicine. 2018. *Assessing and Responding to the Growth of Computer Science Undergraduate Enrollments*. The National Academies Press, Washington, DC. DOI:https://doi.org/10.17226/24926

[51] Zacharoula Papamitsiou, Michail Giannakos, Simon, and Andrew Luxton-Reilly. 2020. Computing education research landscape through an analysis of keywords. In *Proceedings of the ACM Conference on International Computing Education Research*. 102–112. DOI:https://doi.org/10.1145/3372782.3406276

[52] Arnold Pears, Stephen Seidman, Crystal Eney, Päivi Kinnunen, and Lauri Malmi. 2005. Constructing a core literature for computing education research. *ACM SIGCSE Bull.* 37, 4 (2005), 152–161. DOI:https://doi.org/10.1145/1113847.1113893

[53] Arnold Pears, Stephen Seidman, Lauri Malmi, Linda Mannila, Elizabeth Adams, Jens Bennedsen, Marie Devlin, and James Paterson. 2007. A survey of literature on the teaching of introductory programming. In *Working Group Reports on ITiCSE on Innovation and Technology in Computer Science Education (ITiCSE-WGR'07)*. Association for Computing Machinery, New York, NY, 204–223. DOI:https://doi.org/10.1145/1345443.1345441

[54] Manuel A. Perez-Quinones and Elizabeth K. Hawthorne. 2017. Multiple paper types for SIGCSE 2018. *SIGCSE Bull.* 49, 3 (July 2017), 7–7. DOI:https://doi.org/10.1145/3129166.3129172

[55] James Prather, Brett A. Becker, Michelle Craig, Paul Denny, Dastyni Loksa, and Lauren Margulieux. 2020. What do we think we think we are doing? Metacognition and self-regulation in programming. In *Proceedings of the ACM Conference on International Computing Education Research*. 2–13. DOI:https://doi.org/10.1145/3372782.3406263

[56] Justus J. Randolph, George Julnes, Erkki Sutinen, and Steve Lehman. 2008. A methodological review of computer science education research. *J. Inf. Technol. Educ.: Res.* 7 (2008), 135–162. DOI:https://doi.org/10.28945/183

[57] Per Runeson and Martin Höst. 2009. Guidelines for conducting and reporting case study research in software engineering. *Empir. Softw. Eng.* 14, 2 (2009), 131. DOI:https://doi.org/10.1007/s10664-008-9102-8

[58] 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 International Computing Education Research Conference (ICER'19)*. 177–185. DOI:https://doi.org/10.1145/3291279.3339408

[59] Stefan Schmidt. 2009. Shall we really do it again? The powerful concept of replication is neglected in the social sciences. *Rev. Gen. Psychol.* 13, 2 (2009), 90–100. DOI:https://doi.org/10.1037/a0015108

[60] Kenneth F. Schulz, Douglas G. Altman, and David Moher. 2010. CONSORT 2010 statement: Updated guidelines for reporting parallel group randomised trials. *BMJ* 340 (2010). DOI:https://doi.org/10.1136/bmj.c332

[61] Judy Sheard, Simon, Margaret Hamilton, and Jan Lönnberg. 2009. Analysis of research into the teaching and learning of programming. In *Proceedings of the 5th International Workshop on Computing Education Research*. ACM, 93–104. DOI:https://doi.org/10.1145/1584322.1584334

[62] Simon. 2007. A classification of recent Australasian computing education publications. *Comput. Sci. Educ.* 17, 3 (2007), 155–169. DOI:https://doi.org/10.1080/08993400701538021

[63] Simon. 2007. Koli calling comes of age: An analysis. In *Proceedings of the 7th Baltic Sea Conference on Computing Education Research (Koli Calling'07)*. Australian Computer Society, Inc., AUS, 119–126.

[64] Simon. 2009. Ten years of the Australasian Computing Education Conference. In *Proceedings of the 11th Australasian Conference on Computing Education (ACE'09)*. Australian Computer Society, Inc., AUS, 157–164.

[65] Simon. 2020. Twenty-two years of ACE. In *Proceedings of the 22nd Australasian Computing Education Conference (ACE'20)*. Association for Computing Machinery, New York, NY, 203–210. DOI:https://doi.org/10.1145/3373165.3373188

[66] Simon, Angela Carbone, Michael de Raadt, Raymond Lister, Margaret Hamilton, and Judy Sheard. 2008. Classifying computing education papers: Process and results. In *Proceedings of the 4th International Workshop on Computing Education Research*. 161–172. DOI:https://doi.org/10.1145/1404520.1404536

[67] Simon and Judy Sheard. 2020. Twenty-four years of ITiCSE papers. In *Proceedings of the ACM Conference on Innovation and Technology in Computer Science Education*. 5–11. DOI:https://doi.org/10.1145/3341525.3387407

[68] Simon, Judithe Irene Sheard, Angela Carbone, Michael de Raadt, Margaret Hamilton, Raymond Lister, and Errol Thompson. 2008. Eight years of computing education papers at NACCQ. *National Advisory Committee on Computing Qualifications* (2008). http://hdl.handle.net/10453/12628

[69] Claudia Szabo, Nickolas Falkner, Andrew Petersen, Heather Bort, Kathryn Cunningham, Peter Donaldson, Arto Hellas, James Robinson, and Judy Sheard. 2019. Review and use of learning theories within computer science education research: Primer for researchers and practitioners. In *Proceedings of the Working Group Reports on Innovation and Technology in Computer Science Education*. 89–109. DOI:https://doi.org/10.1145/3344429.3372504

[70] Claudia Szabo, Judy Sheard, Andrew Luxton-Reilly, Simon, Brett A. Becker, and Linda Ott. 2019. Fifteen years of introductory programming in schools: A global overview of K–12 initiatives. In *Proceedings of the 19th Koli Calling International Conference on Computing Education Research*. 1–9. DOI:https://doi.org/10.1145/3364510.3364513

[71] Josh Tenenberg. 2014. Asking research questions: Theoretical presuppositions. *Trans. Comput. Educ.* 14, 3 (Sept. 2014). DOI : https://doi.org/10.1145/2644924

[72] David W. Valentine. 2004. CS educational research: A meta-analysis of SIGCSE Technical Symposium Proceedings. *ACM SIGCSE Bull.* 36, 1 (2004), 255–259. DOI : https://doi.org/10.1145/1028174.971391

[73] Arto Vihavainen, Jonne Airaksinen, and Christopher Watson. 2014. A systematic review of approaches for teaching introductory programming and their influence on success. In *Proceedings of the 10th Annual Conference on International Computing Education Research.* 19–26. DOI : https://doi.org/10.1145/2632320.2632349

[74] Jian Zhang and Mark Sherriff. 2020. SIGCSE 2020 recap. *SIGCSE Bull.* 52, 2 (Apr. 2020), 4–7. DOI : https://doi.org/10.1145/3397568.3397570

[75] Stuart Zweben and Betsy Bizot. 2020. 2019 Taulbee Survey. *Comput. Res. News* 32, 5 (2020).