A Gap Analysis of Statistical Data Reporting in K-12 Computing Education Research: Recommendations for Improvement

Monica M. McGill
Knox College, csedresearch.org
Galesburg, IL, USA
monica@csedresearch.org

Adrienne Decker
University at Buffalo
Buffalo, NY, USA
adrienne@buffalo.edu

ABSTRACT

The quality of reporting of experimental results in computing education literature has been previously shown to be less than rigorous. In this study, we first examined research standards set forth by four organizations: American Psychology Association (APA), American Educational Research Association (AERA), What Works Clearinghouse (WWC), and the CONsolidated Standards of Reporting Trials (CONSORT). We selected the most important data standards based on their prominence across all four and the most typical study designs in computing education research. We then examined 76 articles designated as quantitative research studies (K-12) published in ten venues (2012-2018) to determine whether the reporting in these articles met these five standards. Findings indicate that only 48% of these articles report effect size and even fewer (11%) report confidence intervals and levels. We found that reported data did not meet the standard that data should be "reported in a way that the reader could construct effect-size estimates and confidence intervals beyond those supplied in the paper". Additionally, authors used existing instruments less than a quarter of the time (24%) and used instruments with evidence of reliability and validity less than half of the time (39%). We conclude with recommendations for those in the K-12 computing education research community to consider when reporting statistical data in future work so that we can increase the level of rigorous reporting in this growing field.

CCS CONCEPTS

• **Social and professional topics → Computing education**; **Computing education programs**; **Computer science education**.

KEYWORDS

Primary education, secondary education, K-12, statistical data, reporting, effect size, recommendations

ACM Reference Format:
Monica M. McGill and Adrienne Decker. 2020. A Gap Analysis of Statistical Data Reporting in K-12 Computing Education Research: Recommendations for Improvement. In *The 51st ACM Technical Symposium on Computer Science Education (SIGCSE ’20), March 11–14, 2020, Portland, OR, USA.* ACM, New York, NY, USA, Article 4, 7 pages. https://doi.org/10.1145/3328778.3366842

1 INTRODUCTION

Over the years, researchers have examined the computing education research (CER) field as a whole to determine what content areas are being researched (e.g., introductory programming) [20], for models and methods of the research [9], for theoretical underpinnings of the research [21], and for possible missing areas of exploration [19]. Others have been interested in the level of rigor and reporting in computing education research, with efforts to document how research is conducted in computing education and uncovering the areas the field still needs to grow [1, 27, 28, 32]. In addition, there is a general lack of evaluation instruments with evidence of validity and reliability available for use [11, 22].

Within research in K-12 computing education, previous research has shown that studies and interventions are not well described, missing key variables for understanding the populations studied and the intervention its[10, 12, 23]). Researchers have also discovered weaknesses in the way statistics are reported and used, such as under-reporting of effect size [24, 27, 29].

Considering statistical analysis and reporting of results, the American Psychology Association (APA), What Works Clearinghouse (WWC), American Education Research Association (AERA), and CONsolidated Standards Of Reporting Trials (CONSORT), among others, have set recommendations and standards for reporting statistical data [2, 4, 7, 13, 16]. These standards bodies, however, go further than offering recommendations. They actively promote and even require these standards for articles submitted to publication venues that they oversee. Other journals and publication venues have also required that these standards be met for publication. They recognize that an adequate level of reporting of data is needed to compare studies, to replicate and reproduce studies, and to use in meta-analysis [7, 13, 24, 25]. There is little evidence on the state of statistical data reporting in computing education research, with the most recent example focused on a single venue [29] .

As the CSforAll movement continues to grow in the United States and around the world, determining the baseline of statistical reporting in recent studies can enable the creation of recommendations to be used in the field and increase its level of rigor [3, 17, 30]. Thus, we designed a study to compare a subset of reporting standards for statistical data in the aforementioned entities (APA, AERA, WWC, and CONSORT) against K-12 computing education research. The overarching research question for this study was: *What statistical*

SIGCSE ’20, March 11–14, 2020, Portland, OR, USA

ACM ISBN 978-1-4503-6793-6/20/03.
https://doi.org/10.1145/3328778.3366842

data is reported in K-12 computing education research and how does that align with best practices? From this, we developed two research questions that are the focus of this study:

- RQ1: What are commonly recommended standards for reporting statistical data that are important to computing education research?
- RQ2: What are the gaps in reporting statistical data in K-12 computing education research compared to these known standards?

This study focuses on identifying a limited set of standards that could be adapted quickly with maximum impact on the computing education research community. This study is important to evaluators and researchers who are often authors and reviewers of these articles and who may not be familiar with standard practices in reporting data. Not only do guidelines help when writing a paper that includes statistical data, it can also be used by those reviewing these papers for potential acceptance into publication venues.

Finally, as the community starts to adapt practices needed for collecting and analyzing data on a larger scale, without proper data reporting, the ability to replicate, reproduce, and conduct meta-analysis on studies will be limited [7, 24, 25, 29]. Better reporting will benefit the community's awareness of how different sets of published data actually relate and therefore aid in identifying best practices based on empirical evidence [3].

2 STANDARDS FOR REPORTING STATISTICAL DATA IN RESEARCH

To answer *RQ1: What are most commonly recommended standards for reporting statistical data that are important to computing education research?*, we identified a set of significant standards from each of four standards bodies in the field of research standards: American Psychology Association (APA), American Educational Research Association (AERA), What Works Clearinghouse (WWC), and CONsolidated Standards of Reporting Trials (CONSORT) [2, 4, 7, 13, 16]. Although a complete analysis of statistical data guidelines (including methodological considerations and interpretation of the data) across all the sets of standards is outside the scope of this particular analysis, we selected a few standards from each subarea that were most relevant to computing education research based on previous studies and our familiarity with the field.

In this section, we briefly provide a summary of each standard body, along with Table 1 that synthesizes specific standards on reporting statistics across the four sets. This is followed by a more detailed look at previous research addressing this issue in computing education research.

2.1 Across related fields

We provide a brief overview of the four standards bodies that we targeted for inclusion in this analysis, including a summary table (Table 1) that shows several best practices recommended across the four. Although the recommendations in each use slightly different terminology and are organized differently, we carefully considered each and determined whether or not to include it and how to adapt it to the larger set being compared. Specifically, we included:

- Methodological and interpretation guidelines if they are only one-step removed from the actual reporting of data
- Guidelines specifically relevant to quantitative studies
- Guidelines closely related to one-group pre-test/post-test studies
- Guidelines that addressed issues previously reported in computing education research studies

To analyze, one researcher reviewed the identification of individual guidelines that would qualify, added each of these to a list, then began the process of carefully reviewing each to determine if they belonged or if they were similar in nature. For those similar in nature but with different wording, the researcher reworded the guideline to be more generic, encompassing two or more guidelines. The researcher then added these to the table and identified which appeared in which standard body. Once completed, the second researcher reviewed the list for any duplicate entries. The researchers discussed anomalies and addressed these within the table.

2.1.1 American Psychological Association. The 6th edition of the APA Publication Manual (2010) is a style manual that provides a thorough description of *how* to report data stylistically, but only gives limited insight into *what* to report [4]. For example, it presents sample tables on presenting psychometric properties of variables, recommending the variable name, *n* (number of participants), *M* (mean), *SD* (standard deviation), α, range, and skew. It recommends reporting "sufficient information", which is dependent "...on the analytic approach reported" [4, p. 116].

The APA calls out the importance of reporting effect size with its precision both in narrative and in sample tables. The one-degree-of-freedom contrasts table, for example, shows the variables (constructs) at different times (Time 1 and Time 2), with *M* and *SD* for the times being compared, t(34) noting 34 degrees of freedom with the appropriate *t*-values, *p* values, with a 95% Confidence Interval (CI) listing the lower and upper confidence levels, along with Cohen's *d* [4, p. 142].

The APA has guidelines for Journal Article Reporting Standards (JARS) and Meta-Analysis Reporting Standards (MARS). The APA provides a separate book on the reporting of quantitative research in psychology that expands on the Journal Article Reporting Standards (JARS) and Meta-Analysis Reporting Standards (MARS) [8].

2.1.2 American Educational Research Association. The American Educational Research Association (AERA) was founded in 1916 to "...advance knowledge about education, to encourage scholarly inquiry related to education, and to promote the use of research to improve education and serve the public good" [2]. One of the areas of focus for the AERA is to improve quality research practices in education, including improving peer review processes of publication venues. Part of this is the development of statistical data reporting standards for the methodology and the analysis and interpretation of data [13]. Their comprehensive view includes data-driven research to teach, engage, and guide students through better policy investments and to help understand emerging issues and challenge conventional practices [2, 3]. Unlike the APA, the AERA is strictly focused on educational research, particularly at the K-12 level. Their standards identify and promote best practices in high quality research.

Table 1: Essential recommendations relevant across commonly referenced standards.

APA	AERA	WWC	CONSORT	Statistical Reporting Recommendation
				Methodology
X	X	X	X	State statistical analysis conducted and their appropriateness
X	X	X		Provide rationale and evidence of validity of instruments/measurements used (previously validated or ad hoc)
	X	X		Describe any classifications (such as coding of open-ended responses)
X			X	Intended and achieved sample size, power, and precision
X	X	X	X	Describe how missing data/loss of participants is treated
X	X			Provide information concerning problems with statistical assumptions and/or data distributions that could affect validity of findings
X	X			Provide references for less commonly used statistics, used unconventionally or controversially, or statistic is article focus
X				Describe statistical software program, if specialized procedures were used
				For each primary/secondary outcome and for each subgroup
X	X	X		Summary of cases included/deleted from each analysis
X	X	X	X	Subgroup sample sizes (means, SD, other estimates of precision and descriptive statistics) so "reader could construct effect-size estimates and CIs beyond those supplied in paper"
			X	How heterogeneity among participants/subgroups was assessed
X				Number of deliverers; in cases of interventions, include the M, SD, and range of number of individuals/units treated by each
X		X		Whether analysis was "intent to treat, complier average causal effect, other, or multiple ways"
				For Inferential Statistics (e.g., t-tests, F tests, χ^2 tests, effect size and Confidence Intervals (CI) used in null hypothesis significance testing)
X	X	X	X	Sufficient information so readers fully understand/replicate analysis conducted
	X			Information about the "*a priori* type I error rate adopted"
X	X	X		Standard error of the mean
X	X	X		Direction, magnitude, degrees of freedom, and the exact p level
X	X	X	X	Effect sizes (estimate (regression coefficient or difference in means (including odds ratio), p-value and effect size)
X	X	X	X	Effect size precision (Confidence Intervals) were appropriate and levels of confidence
			X	Both absolute and relative effect sizes for binary outcomes
X				Summary statistics for each level of aggregation in multilevel models
X		X		Data results even if no significant effect is found
	X	X		For hypothesis testing, test statistic and its associated significance level
X				For multivariate analysis, include associated variance-covariance (or correlation) matrix/matrices and estimation problems (e.g., failure to converge, bad solutions spaces), anomalous data points
				Interpretation of results
X		X	X	Distinguishing pre-specified analysis (primary and secondary) from exploratory (adjusted analysis)
	X	X		Describe how analysis and presentation of outcomes support claims or conclusions, including when baseline and outcome measures are same and have strong relationship
	X			Describe considerations that are identified during the data analysis
X				Discussion of implications of ancillary analyses for statistical error rates
X	X		X	Present any problems or limitations with data/statistics that could affect validity of findings
	X			Narrative interpretation of index of effect describing it in context to research question(s)

2.1.3 *What Works Clearinghouse.*

Under the (U.S.) National Center for Education Evaluation and Regional Assistance within the Institute of Education Sciences, the What Works Clearinghouse (WWC) represents an effort to synthesize the "best evidence of the effectiveness of education programs, policies, practices and reports" [15]. Since the WWC is a repository housed within the Department of Education, it incorporates many of the critical reporting requirements needed to build well-vetted best practices for K-12 learners. The WWC model calls on qualified researchers and evaluators to review submitted already-published articles by evaluating all aspects of the study against rigorous guidelines [16]. Not only do the articles serve as exemplars for other researchers, they also provide strong data results for determining best practices for curriculum designers, teachers, and policymakers. The WWC handbook is intended for reviewers, so it is less detailed about what exactly to report and more detailed on interpreting what is reported.

To illustrate their level of rigor, the WWC differentiates between random control trials (RCTs) and quasi-experimental designs (QEDs), noting that "RCTs with high attrition, compromised RCTs, and all QEDs are ineligible to receive the highest WWC rating because of uncertainty about intervention and comparison group similarity prior to the introduction of the intervention." [17, p. 14]

2.1.4 *CONSORT.*

The CONsolidated Standards Of Reporting Trials (CONSORT) is designed to "to alleviate the problems arising from inadequate reporting of randomized controlled trials" in an effort to improve the quality of published healthcare research [7]. Initially formed in Canada by researchers recognizing the need to improve data reporting for the medical community, the standards have been endorsed by international medical journals (general and specialized) and editorial organizations. Because of its medical research focus, it focuses only on data related to randomized control trials in the field of medicine, including N-of-1 studies [7]. However, given the advancement of quantitative analysis and formal research processes in the field, it is worth investigating what type of statistical data reporting is required in this type of study design. Though the majority of studies in our dataset are quasi-experimental (one group pre-test post-test) studies, there are some that are randomized control trials. Meta-analysis is a major component in determining best practices in the medical community and reporting standards like these could enable the computing education research community to conduct meta-analysis with higher quality data.

CONSORT provides a two and a half page checklist, with additional information online. This 25 item (with 12 subitems) list serves as a quick reference for researchers to verify that their article meets standards prior to submitting to a journal. It can also be a reference for reviewers when looking at submissions to see that they meet the standards.

2.2 Reporting in Computing Education Research

Statistical data reporting has been identified as an issue in computer science education research due to the abundance of anecdotal evidence reported in articles and experimental designs [6, 14, 27, 32]. Additionally, only 1 in 65 articles (1%) reviewed by Randolph et al. (2018) were determined to have evidence of validity for the evaluation instruments used in the study [27]. In 2019, Margulieux et al. analyzed published papers across three venues for evidence of their use of measurement and concluded that even though there are several standardized measures available with evidence of validity, a majority of studies did not use such measures [22] .

In 2008, Sanders et al. identified key components of data with respect to ethical reporting [28], recommending that authors should share data to maximize the value of each participants' (and each study's) contribution. More recently, Sanders et al. (2019) conducted a study that was solely focused on statistical data (inferential) in computer science education research. Using the definition of adequate described by Randolph et al. [27], they reported that of 270 International Computing Education Research (ICER) conference papers (2005-2018), parametric tests (t-tests) were adequately reported only 37% of the time [29]. With respect to non-parametric tests, chi-squared was adequately reported only 59% of the time and Mann-Whitney-Wilcoxon was adequately reported only 55% of the time. The authors also note in the study that the lack of effect size reporting in these articles was concerning.

Given the above analysis, we carefully considered the most relevant standards that were raised across the four standards bodies and considered each against the needs raised in computing education research (see Table 1). Therefore, instead of including *all* standards related to statistical data analysis recommended by the four standards bodies, we carefully chose the most common for the types of studies most often conducted (quasi-experimental, quantitative studies) [24].

3 METHODOLOGY

To answer *RQ2: What are the gaps in reporting statistical data in K-12 computing education research compared to these known standards?*, we examined these targeted standards (Section 2.2) against K-12 computing education research articles by carefully reviewing 510 articles (2012-2018) across ten venues, including ACM and IEEE publication venues, journals and conference proceedings, and two additional journals and conference proceedings solely focused on computing education research [23] [1]. These articles were specially curated by the https:\csedresearch.org team for this resource center created for the computing education research community.

Based on the above, the most relevant standards promoted across the four standards bodies and our experiences previously analyzing data in articles, we chose the following to investigate in our dataset:

- Methodology
  - How is missing data/loss of participants treated
  - Evidence of validity of evaluation instruments used
  - Describe statistical analysis conducted and appropriateness
- Descriptive/Basic Inferential Statistics
  - For each primary/secondary outcome, group/subgroup sample sizes (means, SD, other estimates of precision so users could construct their own estimates)
- Inferential statistics

[1] ACM International Computing Education Research, Innovation and Technology in Computer Science Education, SIGCSE Technical Symposium on Computer Science Education, Transactions on Computing Education; IEEE Frontiers in Education, Global Engineering Education Conference, Transactions on Education; Journal of Educational Computing Research; Koli Calling; Taylor & Francis' Computer Science Education

- Effect size and effect size precision (Confidence intervals, levels of confidence)
- Report data even if no significant effect is found

Of the original 510 candidate articles, 247 articles were classified as research studies, 47 (19%) of these were quantitative and 114 (46%) were classified as mixed methods, giving a total of 161 (65%) that reported quantitative data. Using this set of papers, we placed them all in order that they were returned by the SQL query (based on their unique identifier in the table) and then took a sampling (every other one on the list) to be reviewed for further analysis. We reviewed 81 papers further to classify their content against the targeted elements by dividing the list equally for review between two raters. The two raters coded the first article concurrently to reach agreement on coding. The subsequent articles were coded independently with discussion between raters about uncertainties. Four articles were removed due to them being mislabelled (qualitative) and one was removed due to it being theory. This left 76 papers for analysis.

4 RESULTS

After analyzing the 76 papers, we found that 44 (58%) were One-Shot Case Study designs and 21 (28%) were One-Group Pretest-Posttest designs. There were 6 (8%) Static-Group Pretest-Posttest studies and 3 (4%) Static-group Comparisons. Of this set, only one (1%) qualified for classification as a Randomized Post-test Only, Control Group Design study.

4.1 Methodological Content

Of the 76 articles analyzed, only 17 (22%) mentioned either how missing data was treated or how loss of participants affected their analysis. The vast majority (79%) failed to mention either, indicating a gap in the understanding of the importance of this measure.

For evaluation instruments/assessment measures, 56 (74%) used ad hoc instruments created by the authors, 17 (22%) used existing instruments, and 1 (1%) study used both. Of these instruments, the majority of studies (45 or 61%) did not report any evidence of validity or reliability. Of the remaining studies, 14 (19%) reported basic data on validity and reliability from previously existing studies, 6 (8%) reported evidence of validity and reliability, 7 (9%) reported evidence of reliability only, and 2 (3%) reported evidence of validity only. This shows that the majority of researchers are creating their own instruments rather than using existing ones, and there is also a failure to show evidence of validity or reliability in these instruments. Some of the studies used various pieces of existing instruments and merged them together into a new instrument but failed to provide evidence of validity of the new instrument.

Of the 75 (out of 76) papers in which it was applicable, only 21 (28%) fully explained and only 12 (16%) partially explained the statistical analysis that was being conducted and how that was relevant to the study. The majority (42 or 56%) did not offer any information about the statistics to be used in conducting the analysis or why they were appropriate for the study.

4.2 Descriptive/Basic Statistics

We found that 52 (69%) of the articles reported the number of participants (*n*) for all groups and subgroups and 15 (20%) reported them partially. Nine (12%) did not adequately report the number of participants, either by not reporting the sizes of the subgroups or failing to report any information about the size of the sample for the study.

Of the 62 articles that reported means, the reporting of mean averages and their variance was evaluated through the entire paper, and no effort was made in this particular high-level analysis to separate means and standard deviation/standard error of mean as reported for descriptive statistics or for inferential statistics. The numbers, therefore, reflect these values in both areas. We found that 45 (73%) reported means for all areas that were evaluated and 3 (5%) partially reported the means. For standard deviation, 29 (50%) of the 58 qualifying articles fully reported and 3 (5%) partially reported standard deviation, leaving nearly half (26 or 45%) that did not report this important measure.

4.3 Inferential Statistics

Of the 64 articles reporting inferential statistics that could report effect size, odds ratio, or regression coefficient, 27 (42%) fully reported this important measure and 4 (6%) partially did. However, nearly half of the articles (33 or 52%), did not. Of the 63 articles that could report the effect size precision (e.g., confidence intervals and levels), only 5 (8%) articles did fully and 2 (3%) did partially, leaving 56 (89%) that did not report this measure at all.

Finally, all four standards bodies call on the importance of reporting data when no significant difference is found. In our collection of 53 articles that had inferential data to report, 39 (53%) fully reported this data (for all analyses) while 3 (4%) partially reported it.

5 DISCUSSION AND RECOMMENDATIONS

Studies like this have been conducted across several fields. For example, in a 2012 study of research articles related to cancer, 35 (43%) reported the amount of missing data according to the suggested guidelines, with the authors concluding that journals should require guidelines like those in STrengthening the Reporting of OBservational studies in Epidemiology (STROBE) [18, 31]. The CONSORT standards are much stronger and have evolved beyond STROBE for the medical field. Similarly, the WWC, AERA, and APA, with even the 7th edition of the APA Publication Manual being available in October 2019, have all evolved, each with the goal to enable high quality research studies in their respective fields.

The availability of the information about the study and methods used to conduct it are key to reproducibility and further advancement of the field. The U.S. National Science Foundation (NSF) released guidelines in 2018 regarding the importance of replication and reproducibility [25]. The ability to reproduce results has proven elusive in closely related fields like social science and psychology [5, 26], which have longer standing established guidelines for reporting. Given the field of computing education is so new, replication and reproducibility is perhaps even more important. Enabling these practices during its foundational roots will help reinforce good research practices as acceptable and expected.

All four standards bodies and the NSF recommend that sufficient information is provided so that readers fully understand the study and can replicate and/or reproduce its analysis. This single basic, yet abstract, standard can be made more concrete by providing recommendations for the data and information closely related to data

Table 2: Recommendations for Reporting Statistical Data in Computing Education Research, Version 1.0.

Recommendations	Examples
In the methodology section	
Describe how missing data and or loss of participants is treated.	"15 students participated in the pre-test. Due to two participants leaving camp early, only 13 students participated in the post-test. Using unique IDs that the participants added to their surveys to the pre- and post-tests (last four digits of their phone number and their middle initial), we removed their pretests from the study."
When using an existing evaluation instrument or assessment measure, provide a brief synopsis of its evidence of reliability and validity and the context of this evidence.	"We used the ABC Instrument to determine student's self-efficacy. ABC Instrument has been previously shown to have Cronbach's α of 0.78 with further evidence for validity presented in [reference]."
When using an ad hoc (your own) instrument or assessment measure, provide evidence of validity and reliability.	"We tested for reliability using Cronbach's α on each of the two constructs, Interest (α=0.73) and Persistence (α=0.81). Our evidence for validity is ...."
For each statistical method to be used, provide an explanation of each statistical analysis to be conducted and its appropriateness.	"It was determined that the *SD* of the two groups was different enough to violate the homogeneity test, $p = .04$; therefore, the non-parametric and more conservative Mann-Whitney test was used to compare groups, $U = 4576, p = .001$."
Descriptive/Basic Inferential Statistics	
For each primary and secondary outcome, report all group and subgroup sample sizes.	"There were 45 students total who participated in the study, 20 in classroom A and 25 in classroom B."
Report means, standard deviation, other estimates of precision for demographic data and for inferential statistical analysis.	"For the pre-test, girls reported lower means in self-efficacy ($M = 2.71$, $SD = 0.45$) than boys ($M = 3.62$, $SD = 0.60$)."
Inferential Statistics	
Report effect size, regression coefficient, or odds ratio.	"Cohen's effect size value (d = .62) suggests a moderate to high significance...."
Report effect size precision (confidence intervals and levels).	"... 95% *CI* [17.2, 43.7]"
Report data even if no significant effect is found.	"There was no statistically significant difference between female students ($M = 4.5$, $SD = 0.36$) and male students ($M = 4.3$, $SD = 0.29$), $t(45) = 1.15, p = .067$, 95% $CI[-19.32, 5.16]$."
For Publication Venues	
Provide space accommodations in publication venues, such as increased page limits.	
Provide guidance where appropriate on statistical data to be reported, with examples of how to report.	
Provide clear guidance to reviewers on what statistical data and how statistical data should be reported in articles under review.	

to report in studies. To do this for the computing education research field, we conclude this study with a set of recommendations, *Recommendations for Reporting Statistical Data in Computing Education Research, Version 1.0* in Table 2. In addition to recommendations for researchers and reviewers, we provide three recommendations for computing education research publication venues to consider.

6 CONCLUSION

Without adequate reporting, decisions about best practices affecting thousands of children and costing millions of dollars may be made on inadequate data, which is why the standards bodies are so interested in improving data reporting [15]. Their recommendations have been developed so that researchers not involved in a study are able to fully understand them for comparing, for replicating and reproducing, and for conducting meta-analysis.

An analysis like this could become quite extensive, particularly if we examined more standard bodies and their recommendations. Our goal in this study is to provide basic recommendations that will not only give the community food for thought, but also provide additional momentum to move the community closer to publishing articles and papers that further meet high quality research practices.

Though we have labelled this set of recommendations as version 1.0, as the community starts to realize the importance of these recommendations and to adopt them, we will continue to build on this work so that we can help the community reach even higher standards of data reporting comparable to other fields.

ACKNOWLEDGMENTS

This material is based upon work supported by the U.S. National Science Foundation uder Grant Nos. 1625335, 1649671, 1745199, 1757402 and 1933671. We also extend our thanks to our csedresearch.org team (https://csedresearch.org/meet-the-team) and our RPPforCS colleagues at SageFox and CSforAll for their continued support.

REFERENCES

[1] 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)*. ACM, New York, NY, USA, 120–125. https://doi.org/10.1145/2839509.2844601

[2] American Educational Research Association. 2019. American Educational Research Association. Retrieved August 21, 2019 from https://www.aera.net/

[3] American Educational Research Association. 2019. When Researchers Have Access to Data, Students Succeed. Retrieved August 21, 2019 from https://www.aera.net/Portals/38/docs/Policy_and_Programs/DQC%20-%20Research%20v2_digital.pdf?ver=2017-10-27-150504-387

[4] American Psychological Association. 2010. *Publication manual of the American Psychological Association (6th ed.)*. American Psychological Association: Washington, DC, USA.

[5] Colin F Camerer, Anna Dreber, Felix Holzmeister, Teck-Hua Ho, Jürgen Huber, Magnus Johannesson, Michael Kirchler, Gideon Nave, Brian A. Nosek, Thomas Pfeiffer, et al. 2018. Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nature Human Behaviour* 2, 9 (2018), 637.

[6] Mike Clancy, John Stasko, Mark Guzdial, Sally Fincher, and Nell Dale. 2001. Models and areas for CS education research. *Computer Science Education* 11, 4 (2001), 323–341.

[7] CONSORT Group. 2010. CONSORT 2010 checklist of information to include when reporting a randomised trial. Retrieved August 21, 2019 from http://www.consort-statement.org/download/Media/Default/Downloads/CONSORT%202010%20Checklist.doc

[8] Harris Cooper. 2018. *Reporting Quantitative Research in Psychology: How to Meet APA Style Journal Article Reporting Standards*. American Psychological Association.

[9] Mats Daniels and Arnold Pears. 2012. Models and Methods for Computing Education Research. In *Proceedings of the Fourteenth Australasian Computing Education Conference - Volume 123 (ACE '12)*. Australian Computer Society, Inc., Darlinghurst, Australia, Australia, 95–102. http://dl.acm.org/citation.cfm?id=2483716.2483728

[10] Adrienne Decker and Monica M. McGill. 2017. Pre-College Computing Outreach Research: Towards Improving the Practice. In *Proceedings of the 2017 ACM SIGCSE Technical Symposium on Computer Science Education (SIGCSE '17)*. ACM, New York, NY, USA, 153–158. https://doi.org/10.1145/3017680.3017744

[11] 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)*. ACM, New York, NY, USA, 558–564. https://doi.org/10.1145/3287324.3287393

[12] Adrienne Decker, Monica M. McGill, and Amber Settle. 2016. Towards a Common Framework for Evaluating Computing Outreach Activities. In *Proceedings of the 47th ACM Technical Symposium on Computing Science Education (SIGCSE '16)*. ACM, New York, NY, USA, 627–632. https://doi.org/10.1145/2839509.2844567

[13] Richard P. Duran, Margaret A. Eisenhart, Frederick D. Erickson, Carl A. Grant, Judith L. Green, Larry V. Hedges, and B.L. Schneider. 2006. Standards for reporting on empirical social science research in AERA publications: American Educational Research Association. *Educational Researcher* 35, 6 (2006), 33–40.

[14] Christian Holmboe, Linda McIver, and Carlisle E. George. 2001. Research Agenda for Computer Science Education. In *PPIG*, Vol. 13.

[15] Institute of Education Sciences. [n. d.]. About the WWC. Retrieved August 26, 2019 from https://ies.ed.gove/ncee/aboutus/

[16] Institute of Education Sciences. [n. d.]. What Works Clearinghouse: Reporting Guide for Study Authors: Group Design Studies. Retrieved August 21, 2019 from https://ies.ed.gov/ncee/wwc/Docs/ReferenceResources/wwc_gd_guide_022218.pdf

[17] Institute of Education Sciences. [n. d.]. What Works Clearinghouse Standards Handbook Version 4.0. Retrieved August 21, 2019 from https://ies.ed.gov/ncee/wwc/Docs/referenceresources/wwc_standards_handbook_v4.pdf

[18] Amalia Karahalios, Laura Baglietto, John B. Carlin, Dallas R. English, and Julie A. Simpson. 2012. A review of the reporting and handling of missing data in cohort studies with repeated assessment of exposure measures. *BMC medical research methodology* 12, 1 (2012), 96.

[19] Päivi Kinnunen and Beth Simon. 2010. Building Theory About Computing Education Phenomena: A Discussion of Grounded Theory. In *Proceedings of the 10th Koli Calling International Conference on Computing Education Research (Koli Calling '10)*. ACM, New York, NY, USA, 37–42. https://doi.org/10.1145/1930464.1930469

[20] Andrew Luxton-Reilly, Simon, 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 Companion of the 23rd Annual ACM Conference on Innovation and Technology in Computer Science Education (ITiCSE 2018 Companion)*. ACM, New York, NY, USA, 55–106. https://doi.org/10.1145/3293881.3295779

[21] 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 Tenth Annual Conference on International Computing Education Research (ICER '14)*. ACM, New York, NY, USA, 27–34. https://doi.org/10.1145/2632320.2632358

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

[23] 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.

[24] Monica M McGill, Tom McKlin, and Errol Kaylor. 2019. Defining What Empirically Works Best: Dynamic Generation of Meta-Analysis for Computer Science Education. In *Proceedings of the 2019 ACM Conference on International Computing Education Research*. ACM, 199–207.

[25] National Science Foundation and Institute for Education Science. 2018. Companion Guidelines on Replication & Reproducibility in Education Research: A Supplement to the Common Guidelines for Education Research and Development. https://ies.ed.gov/pdf/CompanionGuidelinesReplicationReproducibility.pdf. Accessed: 2019-03-01.

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

[27] J. Randolph, J. G. Julnes, S. Erkki, and S. Lehman. 2008. A methodological review of Computer Science Education research. 7 (2008), 135âĂŞ162.

[28] Kate Sanders, Brad Richards, Jan Erik Moström, Vicki Almstrum, Stephen Edwards, Sally Fincher, Kat Gunion, Mark Hall, Brian Hanks, Stephen Lonergan, Robert McCartney, Briana Morrison, Jaime Spacco, and Lynda Thomas. 2008. DCER: Sharing Empirical Computer Science Education Data. In *Proceedings of the Fourth International Workshop on Computing Education Research (ICER '08)*. ACM, New York, NY, USA, 137–148. https://doi.org/10.1145/1404520.1404534

[29] Kate Sanders, Judy Sheard, Brett A. Becker, Anna Eckerdal, Sally Hamouda, and Simon. 2019. Inferential Statistics in Computing Education Research: A Methodological Review. In *Proceedings of the 2019 ACM Conference on International Computing Education Research (ICER '19)*. ACM, New York, NY, USA, 177–185. https://doi.org/10.1145/3291279.3339408

[30] Megan Smith. 2016. Computer Science for All. https://obamawhitehouse.archives.gov/blog/2016/01/30/computer-science-all

[31] STROBE. 2012. STrengthening the Reporting of OBservational studies in Epidemiology. Retrieved August 24, 2019 from https://strobe-statement.org/index.php?id=strobe-home

[32] David W. Valentine. 2004. CS Educational Research: A Meta-analysis of SIGCSE Technical Symposium Proceedings. In *Proceedings of the 35th SIGCSE Technical Symposium on Computer Science Education (SIGCSE '04)*. ACM, New York, NY, USA, 255–259. https://doi.org/10.1145/971300.971391