Inferential Statistics in Computing Education Research: A Methodological Review

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

Judy Sheard
Monash University
Melbourne, Australia, Australia
judy.sheard@monash.edu

Brett A. Becker
University College Dublin
Dublin, Ireland
brett.becker@ucd.ie

Anna Eckerdal
Uppsala University
Uppsala, Sweden
Anna.Eckerdal@it.uu.se

Sally Hamouda
Rhode Island College
Providence, RI, United States
shamouda@ric.edu

Simon
University of Newcastle
Newcastle, Australia
simon@newcastle.edu.au

ABSTRACT

The goal of most computing education research is to effect positive change in how computing is taught and learned. Statistical techniques are one important tool for achieving this goal. In this paper we report on an analysis of ICER papers that use inferential statistics. We present the most commonly used techniques; an overview of the techniques the ICER community has used over its first 14 years of papers, grouped according to the purpose of the technique; and a detailed analysis of three of the most commonly used techniques (t-test, chi-squared test, and Mann-Whitney-Wilcoxon). We identify common flaws in reporting and give examples of papers where statistics are reported well. In sum, the paper draws a picture of the use of inferential statistics by the ICER community. This picture is intended to help orient researchers who are new to the use of statistics in computing education research and to encourage reflection by the ICER community on how it uses statistics and how it can improve that use.

KEYWORDS

alphabet soup, empirical research, methodology, reporting, statistical significance, statistics

ACM Reference Format:
Kate Sanders, Judy Sheard, Brett A. Becker, Anna Eckerdal, Sally Hamouda, and Simon. 2019. Inferential Statistics in Computing Education Research: A Methodological Review. In *International Computing Education Research Conference (ICER '19), August 12–14, 2019, Toronto, ON, Canada,* Robert McCartney and Anthony Robins (Eds.). ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3291279.3339408

1 INTRODUCTION

This paper presents a methodological review. While literature reviews look at outcomes (for example, what the literature says about teaching CS1), methodological reviews look at our research practice: how we perform and report our research. Here, we examine how inferential statistics are used in computing education research.

Computing education research has been called a 'trading zone': we borrow theories and methodologies from a number of other fields, including education, social science, and more [8]. Statistics is one of the tools we share with other disciplines. Descriptive statistics – mean, median, mode, standard deviation, and so on – are familiar to most, but inferential statistical techniques are less well known. Computing education researchers, even those who have used statistics in computing research, may have no formal background in the use of statistics for education research. An understanding of this area is important not only for conducting quantitative research, but also for reading and reviewing papers that report on such research, and for questions of inter-rater reliability in qualitative research. Understanding these techniques, and being aware of how to present results effectively, is important for much of the community's research. Our research questions are:

RQ1 How often do computing education research papers include statistical analysis beyond descriptive statistics?
RQ2 What are the most frequently used inferential statistical techniques?
RQ3 Overall, what inferential statistical techniques are used?
RQ4 What is the quality of reporting for some of the more frequently used techniques?

We base our picture of the use of statistics on an analysis of all ICER papers to date (2005-2018). While the overall picture is not encouraging, in conducting this analysis we have come across good examples of reporting as well as less satisfactory ones. We present the results of our analysis, along with some pitfalls to avoid and some examples of good reporting. We also provide some recommendations for computing educators who intend to report statistical results in the future.

2 RELATED WORK

2.1 Methodological Reviews

Two early methodological reviews of computing education research both called for improvements in methodology, such as increasing the number of papers that "made any attempt at assessing the 'treatment' with some scientific analysis" [30, p. 256], providing explicit research questions, including adequate consideration of

ICER '19, August 12–14, 2019, Toronto, ON, Canada
© 2019 Association for Computing Machinery.
ACM ISBN 978-1-4503-6185-9/19/08…$15.00
https://doi.org/10.1145/3291279.3339408

related work, and improving experimental design [21]. These were based respectively on analyses of SIGCSE papers from 1984 to 2003 and Koli Calling papers from 2000 to 2004, and neither addressed the use of statistics.

Inspired by Randolph [20], we classify our review according to Cooper's Taxonomy of Literature Reviews [5], which presents six characteristics, *focus, goal, perspective, coverage, organisation*, and *audience*, each divided into several categories. Our review has a *focus* on 'research methods'. It has two *goals*, to 'integrate and generalise findings' and to 'critically analyse previous research [and] identify central issues' [20, p. 3]. The analysis is mainly quantitative and thus the *perspective* is 'neutral representation'. In terms of *coverage*, our work falls into the category 'purposive sample' where 'central or pivotal articles in a field' (here, ICER papers) are selected. Finally our intended *audience* is 'specialised scholars' (computing education researchers), and the review is *organised* 'methodologically', that is 'as in an empirical paper (i.e., introduction, method, results, and discussion)' [20, p. 4].

Randolph's 2007 thesis [18, 19, 22] embodies an extensive methodological review of computing education research, including the use of statistics. It was based on a stratified proportional random sample of 352 articles published from 2000 to 2005 in a broad range of computing education conferences and journals.

Randolph's analysis of 'statistical practices' (one of the nine research questions in his thesis) is most relevant for our purposes here. Of the 352 articles in his data set, 123 were found to involve human participant data, to contain more than anecdotal evidence, and to be at least partly quantitative. These 123 articles (the 'statistics subset') form the basis for this portion of his analysis; the use of statistics in other articles, such as those not involving data from human participants, was not considered.

Of the 123 articles in Randolph's statistics subset, 44 involved inferential statistics. He assigned each of the 44 papers to one or more of five categories, according to which type(s) of inferential statistics the article included: parametric analysis, non-parametric analysis, correlational analysis, small sample analysis, and multivariate analysis. Specific inferential statistics techniques were not identified.

Randolph concluded that relatively few of the inferential statistics papers he examined reported sufficient information. "Areas of concern include reporting a measure of centrality *and* dispersion for parametric analyses, reporting sample sizes and correlation or covariance matrices for correlational analyses, and summarising raw data when non-parametric analyses are used" [18, p. 146].

While Randolph addressed quantitative research in computing education, López et al. [12] provide a guide for quantitative research in education in general, describing why and how to perform a study and how to analyse the data and report the results.

Several classification schemes have been developed to describe computing education as a research area and to enable improvements in the field. An early classification of computing education publications was applied by Simon and colleagues to papers from ICER as well as other conferences [23–26]. Sheard et al. [23] used Simon's classification to analyse research papers about programming education drawn from several venues, including ICER. They found that of the 41 ICER papers they examined (69% of the ICER papers published from 2005 to 2008), 27 involved quantitative analysis, and 22 (81%) of those made use of statistical techniques beyond descriptive statistics. Malmi et al. [13] extended this analysis and applied it to papers from the first five years of ICER.

While these researchers aimed broadly at developing a classification for "exploring how research in CER is being carried out" [13, p. 9], our analysis narrows the focus to quantitative research, aiming to describe and critically analyse methodology and to identify central issues. Our work is also more limited in that we focus exclusively on one conference, ICER. This choice is supported by Sheard et al. [23]'s finding that ICER has a significantly higher proportion of papers with inferential statistics than other conferences. We extend earlier work in three ways: by including ICER papers regardless of topic, by examining ten additional years of ICER papers (2009-2018), and by adding a more detailed evaluation of three specific and common inferential statistics techniques.

2.2 Standards for Reporting Statistical Research

The American Psychological Association's *Publication Manual* [2] (the *APA Manual*) is used by many writers and editors in the social and behavioural sciences and is the required style for many academic journals. Sections 2 and 4 detail what information should be included when reporting statistical results. The basic principle is to "include sufficient information to help the reader fully understand the analyses conducted and possible alternative explanations for the outcomes of those analyses" [2, p. 33].

Specific requirements follow from this principle: reports must include 'complete reporting of all tested hypotheses', the value of the test statistic (to two decimal places), the degrees of freedom, the exact *p*-value, the effect size, and an adequate description of the data, including the sample size and 'usually at least' the per-cell sample sizes, the observed cell means, and either the cell standard deviations or the pooled within-cell variance [2, pp. 28, 33–34]. For ordinal data, the cell mean should be replaced by the median and for nominal data it should be replaced by a count of the items in the cell. In addition, it is important to take into account the underlying assumptions and limitations of the statistics used: "The methods used must support their analytic burdens, including robustness to violations of the assumptions that underlie them, and they must provide clear, unequivocal insights into the data" [2, p. 33].

Like the APA Manual, the American Educational Research Association's (AERA) standards also require reporting of the test statistic, the significance level, and the effect size [1]. These standards also explicitly note that it is common for non-quantitative data to be transformed into quantitative data for analysis. In computing education research this will be seen, for example, with student software designs and interviews with instructors. Careful reporting of the data transformations and an explanation of their appropriateness is essential because the "validity of empirical studies depends, in part, on the claim that classifications and measurements [i.e., these transformations] preserve important characteristics of the phenomena they represent" [1, Section 4].

3 METHOD

Like that of Randolph [18], our methodological review uses content analysis to "analyse the research practices reported in a body of academic articles" [18, p. 1]. The body of academic articles analysed for this study is the ICER papers from 2005 to 2018, chosen because of ICER's emphasis on empirical research.

Our content analysis is in line with Neuendorf's *Integrative Model of Content Analysis* [16], which Randolph et al. [22, p. 40] describe in the following steps: (a) developing a theory and rationale, (b) conceptualising variables, (c) operationalising measures, (d) developing a coding form and coding book, (e) sampling, (f) training and determining pilot reliabilities, (g) coding, (h) calculating final reliabilities, and (i) analysing and reporting data.

Our method includes these steps, but in an iterative process where most steps were revisited as more data was analysed and the model developed, and not necessarily in the order suggested. For example, throughout the analysis we continued to discuss the variables and their interpretation, since there is no consensus on terminology in the community. As a result, it was sometimes necessary to recode data. Below we give more details of this process.

Analysing the whole data set. We piloted an initial set of categories on the 2005 and 2006 papers. We coded several features related to each paper in our data set; those relevant to this paper include *use of statistics* and *reported statistical methods.* After each researcher had tagged the papers individually, we discussed the results as a group and modified the possible values for the features.

Once we had agreed on the values for the first feature, *use of statistics*, coding was deductive. There are four possible values:

None: No data collected or no numbers used to describe the data.

Basic numerical data: Counts, sums, percentages.

Descriptive: At least one of the following descriptions of the data set: minimum value, maximum value, mean, median, mode, standard deviation, shape of curve (normal, skewed, bimodal, etc.).

Beyond: Inferential statistics and/or techniques such as factor analysis, machine learning, etc. (Note: for convenience we use the term 'inferential statistics' for all of these.)

Where two or more of these categories applied, the most comprehensive one was used; for example, a paper that included both basic numerical data and descriptive elements was tagged as Descriptive.

Papers that use the term 'significant' (in reference to some result, value, etc.) without providing details of any significance test were classified as Descriptive. Papers that include a statistical analysis of inter-rater reliability were considered to use inferential statistics.

The second feature, *reported statistical methods*, applied only to papers in the Beyond category, and, for each paper, we listed the inferential statistics techniques it used. This part of the analysis was necessarily open-ended and inductive, as previously unused techniques continued to be found throughout the analysis.

Once the codes were agreed, the papers were partitioned into sets, each of which was classified by two researchers. Our focus during this process was on reconciliation of differences, and we found that most differences appeared to result from oversights or from simple errors of selection, rather than from genuine differences in perceptions of the papers themselves.

To check the reliability of our deductive analysis (*use of statistics*), we used the 2018 papers (28 of 271, or 10%) as a reliability sample, not having yet tagged these papers. All six researchers independently tagged each of these papers as None, Basic numerical data, Descriptive, or Beyond, and the inter-rater reliability was measured using the Fleiss-Davies kappa [7], a chance-corrected measure for the situation in which all of the classifiers classify all of the items. This was calculated using a purpose-written script acting on data in a spreadsheet, resulting in $\kappa = 0.69$, which is at the high end of the generally agreed 'fair to good' range [3].

To check the reliability of our inductive analysis (*reported statistical methods*), once the pairs had completed and reconciled their analyses a single researcher reviewed all of the results. It was not possible to apply a formal measure of agreement to these results as they are not classifications selected from a fixed set.

Tagging specific inferential statistics techniques. After identifying the most commonly used techniques, we selected three for further analysis: chi-squared tests, t-tests, and Mann-Whitney-Wilcoxon. The chi-squared group of papers includes all papers that mention using a chi-squared test, except for those in which the tests are part of a broader strategy such as structural equation modelling. The Mann-Whitney-Wilcoxon group of papers includes all the papers that use a Mann-Whitney test, a Wilcoxon rank-sum test, or a Wilcoxon signed-rank test, and the t-test group of papers includes all the papers that use any form of t-test.

We then performed a deductive analysis, using a set of tags based on the literature. Utilising the APA Manual [2], we looked for the following: a hypothesis, an *alpha*-value, whether the test is one-tailed or two-tailed (if applicable), whether the test statistic value and exact *p*-value are reported for significant results, and whether the test statistic and descriptive values are reported with an appropriate level of precision. We added another tag that was not stated explicitly, but is consistent with the APA Manual's principle of including sufficient information to help the reader understand: stating the precise name of the test. We also checked whether papers explicitly mention the software used for the statistical calculations.

For comparison with Randolph's results, we considered whether the tests are 'adequately reported' by his definition: essentially, whether sufficient descriptive statistics are provided. For parametric tests such as t-tests, he requires papers to include either cell means and cell sizes, or degrees of freedom and either mean cell variances or mean square error. For non-parametric tests, such as chi-squared and Mann-Whitney-Wilcoxon, his requirement, 'a summary of the raw data', requires some interpretation. For chi-squared tests we took this to mean that the paper reports number of columns, number of rows, and the number of observations in each cell. We considered the requirement satisfied if a paper gives the contingency table, or information from which the contingency table can be created. For Mann-Whitney-Wilcoxon, we took 'summary of the raw data' to mean the sample size, group (cell) size, and median.

Once again, we divided into pairs (different pairs from before), each of which took one of the three techniques and coded the papers using that technique according to the agreed tags.

4 RESULTS

4.1 RQ1: Inferential Statistics Papers

As shown in Table 1, just over half of the 270 ICER papers (51%) use inferential statistics. Somewhat surprisingly, more than a quarter

Table 1: Numbers of papers using various levels of statistics

Type of quantitative analysis	Papers	%
none	76	28
basic numerical data	30	11
descriptive statistics	26	10
beyond descriptive statistics	138	51
TOTAL	270	

Table 2: ICER's most frequently used inferential techniques

Technique	Papers
correlation (different types)	53
t-test (different types)	33
Mann-Whitney, Wilcoxon rank sum, and Wilcoxon signed-rank test	32
chi-squared test (different types)	31
ANOVA (different types)	27
regression (different types)	28
Cronbach's alpha	13
factor analysis	12
Kruskal Wallis	11

(28%) use no numbers to describe their data – in some cases because they have no data.

4.2 RQ2: Most Frequent Techniques

The most frequently used inferential statistical techniques in ICER papers are shown in Table 2. These techniques will be placed in context in the following sections, but they are listed here, so that researchers who are new to the use of statistics in educational research can begin by familiarising themselves with these most frequently used techniques.

4.3 RQ3: All Inferential Statistics Techniques

ICER papers use many different inferential tests. For clarity, we group these tests according to their purpose (for example, finding differences, finding relationships, making predictions). Within each purpose we further group the tests according to the scale of the data, which can be nominal (non-ordered categories, such as whether a student drops out), ordinal (ordered categories, such as a student's response to a Likert question), or interval (continuous variables, such as a student's mark on the final exam).

We examine tests for difference, relationships, predictions, and other purposes one by one in the next four subsections.

4.4 RQ3: Tests for Difference

We consider here tests for difference between groups at the interval and ordinal level scales.

4.4.1 Interval Level Data. Table 3 summarises the tests for difference using interval level data (parametric tests) in ICER papers. These tests look for a difference between the means of two or more groups.

If there are two groups, a *t-test* is typically used. For example, in computing education, a t-test has been used to look for a difference in size between the subgraphs of collaboration networks for computing education conferences and for computer science conferences [14]. There are different types of t-test depending on whether the groups are independent or related (paired) samples. A variation is the one-sample t-test, which is used to determine whether a sample mean is statistically different from a known or hypothesised population mean. The most commonly used t-test in our data set is the independent-samples t-test, with fewer examples of the paired-samples or one-sample tests. One paper reported use of Welch's t-test, an adaption of the t-test that is more reliable when the samples have unequal variance or unequal sample sizes. Given the different types of t-test it is concerning that in 14 (42%) of the cases the exact type of t-test was not specified.

For more than two groups, an *Analysis of Variance (ANOVA)* is typically used to test for a difference between their means. For example, in computing education an ANOVA has been used to determine whether there are significant differences in the number of compilation events per Blackbox user depending on the user's country [11]. There are a number of variations on the ANOVA test depending on how many variables (factors) are tested (one-way ANOVA, two-way ANOVA, etc.) or whether the groups are independent or related samples (repeated-measures). Another variation is the Analysis of Covariance (ANCOVA), which controls for the effects of variables that are not of primary interest (covariates). More advanced variations are the Multivariate Analysis of Variance (MANOVA), which has two or more dependent variables, and the Multivariate Analysis of Covariance (MANCOVA).

The form used most commonly in our data set is the one-way independent-samples ANOVA, with only a couple of examples of two-way and repeated-measures ANOVAs. Three papers reported use of ANCOVA, and two each of MANOVA and MANCOVA. As with t-tests, it is concerning that in 13 (52%) of the cases the type of ANOVA was not specified.

While an ANOVA test can show that there is a difference between the means of a number of groups, it does not show which groups differ from the others. Post-hoc tests can be used for this purpose. For example, in computing education, a Tukey HSD test has been used by Jadud and Dorn [11] (mentioned above as using ANOVA) to determine which countries' users were significantly more active than others. We found 12 examples of post-hoc tests. The Tukey HSD test was the most common, with one example each of Bonferroni, Conover, Dunnett's and post-hoc z-tests.

The results of multivariate tests can be difficult to interpret. A number of test statistics can be used to indicate which effects contribute most to a MANOVA or MANCOVA model. We found two cases where Pillai's trace was used to interpret a MANCOVA.

When multiple tests are performed on a single set of data, there is an increased likelihood of one of the tests being significant purely by chance. A common approach to dealing with this multiple test problem is to use a correction technique to calculate a lower *alpha*-value for testing significance. We found four cases where the Bonferroni correction was used and one case each where the more powerful Benjamini-Hochberg and Holm procedures were used.

Table 3: Tests for difference: interval level data

Name of test	Number of groups	#
one-sample t-test	one	2
independent-samples t-test	two: independent samples	10
Welch two sample t-test	two: independent samples	1
paired-samples t-test	two: related samples	6
t-test (type not specified)	two	14
one-way ANOVA	multiple: one factor, independent samples	7
two-way ANOVA	multiple: two factor, independent samples	2
repeated-measures ANOVA	multiple: related samples	2
ANOVA (type not specified)	multiple	13
ANCOVA	multiple: one factor, type-III, independent samples	1
ANCOVA (type not specified)	multiple	2
MANOVA (type not specified)	multiple	2
MANCOVA (type not specified)	multiple	2

4.4.2 Ordinal Level Data. Table 4 summarises the non-parametric tests for difference (tests that use ordinal level data or data that otherwise violate the assumptions for parametric tests).

Like the parametric tests for difference, the non-parametric tests also apply to cases involving either two or more than two groups. The typical tests for statistical difference between two groups are the Mann-Whitney U test (also known as the Wilcoxon rank-sum test) for independent samples and the Wilcoxon signed-rank test for related samples. For example, in computing education, the Mann-Whitney U test has been used to investigate differences in the benefit received from spatial-skills training by programming students in two socio-economic groups [6].

For testing for difference between more than two groups, the Kruskal Wallis test (for independent samples) was most frequently used. In computing education, Kruskal Wallis was used in the spatial skills study mentioned above, when considering the effect of spatial-skills training on students in three different race/ethnic groups [6]. We also found one case of Friedman's test for related samples. Similarly to the parametric t-tests, the independent samples tests were more frequently used than the related samples test.

Overall we found 45 cases of the use of non-parametric (ordinal data) tests for difference, fewer than the 64 cases of the use of parametric (interval data) tests.

4.5 RQ3: Tests for Relationship

4.5.1 Nominal Level Data. The typical tests for nominal level data are for comparing the proportions of values of a variable to a theoretical distribution (test for goodness of fit) or to proportions of values of other variables (test for independence of variables). Table 5 summarises these tests.

The most common test used was chi-squared, and in one case Pearson's chi-squared was specified. These were used to compare proportions of two or more variables. In one example, a chi-squared test was used to compare the types of questions asked in a closed lab by pair programmers to those asked by students programming on their own [9].

Table 4: Tests for difference: ordinal level data

Name of test	Number of groups	#
Mann-Whitney U test	two: independent samples	11
Wilcoxon rank-sum test	two: independent samples	9
Wilcoxon signed-rank test	two: related samples	11
Wilcoxon (unspecified)	two	1
Kruskal Wallis	multiple: independent samples	11
Friedman's test	multiple: related samples	1
Kendall's coefficient of concordance	multiple: related samples	1

Table 5: Tests for relationship: nominal level data

Name of test	#
Fisher's exact test	2
two proportion z-test	2
chi-squared (Pearson)	31
McNemar-Bowker test	1

Table 6: Tests for relationship: ordinal and interval level data

Name of test	#
Pearson product-moment correlation	17
Spearman's rank-order correlation	9
Kendall's rank correlation	1
correlation (unspecified)	26

There were two cases of Fisher's exact test, which can be used in the place of a chi-squared test for 2×2 contingency tables, especially in cases of small samples. There were also two cases of the two proportion z-test, which is identical to the chi-squared test except that the standard normal deviation is estimated.

We found one example of a McNemar-Bowker test, which is applied to 2x2 contingency tables to compare paired nominal data.

4.5.2 Interval and Ordinal Level Data. The strength of the relationship between two variables at the interval or ordinal level is measured by a correlation test. For example, in computing education, Sirkiä and Sorva [28] found a correlation between assignment scores and use of program visualisation.

Table 6 summarises the correlation tests that we observed. The most common correlation test we found was Pearson's product-moment correlation, which is used for interval level data. Spearman's rank-order correlation was the most commonly used test for ordinal level data. We found one case of Kendall's rank correlation, which is used in small samples or when there are many values with the same score. However, in almost half (49%) of the cases the type of correlation test was not specified.

4.5.3 Factor Analysis. Factor analysis is used to find latent variables (factors) underlying a correlation matrix. There are two forms: exploratory and confirmatory. Exploratory factor analysis (EFA) is conducted to discover the latent variables, whereas confirmatory factor analysis (CFA) is used test a theory or hypothesis about the latent variables expected to be found. We found both types of factor analysis but the form was often not specified. We found examples of several common techniques: principal axis factoring, generalised least squares, and maximum likelihood.

Bartlett's test of sphericity and the Kaiser-Meyer-Olkin (KMO) test are intended to determine the suitability of the correlation matrix for EFA. We found few examples of these tests, and they were used in less than half of the factor analyses that appeared to be exploratory.

Associated with CFA are tests to determine how well a model fits (comparative fit index (CFI) and Tucker-Lewis index (TLI)). We found three cases where these were used.

We also found examples of principal component analysis, not considered a factor analysis but closely related. This technique is used in exploratory data analysis and for making predictive models. Finally, we found several examples of path analysis and structural equation modelling, more advanced techniques involving analysis of relationships.

4.6 RQ3: Tests for Prediction

Regression tests are used to determine the effect of two or more variables (predictors) on dependent variables. For example, in computing education, regression has been used to predict the satisfaction of students with their teams, based on the average 'social sensitivity' score of the team members [4].

There are different types of regression tests and different ways that they can be conducted. In our study the most common type was linear regression and in a couple of cases these were conducted as stepwise regressions. There were two cases of logistic regression and single cases of multivariate, hierarchical linear modelling, and ordinal regression. However, in 30% of cases the type of regression was not specified.

4.7 RQ3: Tests for Other Purposes

4.7.1 Tests on Distributions. For parametric tests there are certain assumptions that should be met about the characteristics of the samples and the populations that the samples are drawn from. We found the most common test used to determine if a sample was normally distributed was the Shapiro-Wilk test (10 cases). Levene's test for equality of variances was used to test for homogeneity of variances between groups (4 cases). We found only one example of a test for skewness and kurtosis matching a normal distribution (Jarque-Bera test) and one example of a test for multi-modality (Hartigan's dip test).

4.7.2 Effect Size. Effect size is a quantitative measure of the magnitude of a statistical difference or relationship and gives an indication of whether the relationship is meaningful. Simply put, an effect size refers to the magnitude of a result, and there are several methods to calculate the size of an effect [31]. We found five cases where Cohen's d was used to measure the effect size of a comparison test and one case where Hedges' g (preferable to Cohen's d for small samples) was used. We found single examples of omega-squared and eta-squared tests used to measure effect sizes. This is concerning as reporting effect size is extremely important – A statistically significant change is of little use if the change doesn't have an effect size indicating that it will make a meaningful difference in the context being examined. Additionally, APA and AERA both require reporting effect sizes as discussed in Section 2.2.

4.7.3 Inter-Rater Reliability. Inter-rater reliability statistics are used to measure the degree to which data are analysed consistently by two or more raters. We found a total of 14 references to various inter-rater reliability measures, including Cohen's kappa, Fleiss-Davies kappa, Krippendorff's alpha, a one-way random intra-class coefficient of agreement, and some unnamed measures. As an example of use, in this paper we have used the Fleiss-Davies kappa to evaluate the inter-rater reliability of coding for the *use of statistics* feature (see section 3).

4.7.4 Reliability. Cronbach's alpha is a measure of the reliability of a test of skills, ability, personality, etc. For example, in computing education, Cronbach's alpha has been used to analyse a survey being developed to measure the cognitive load of a computing/programming intervention [15]. In total, we found 13 papers that used Cronbach's alpha and two that used unspecified reliability tests.

4.8 RQ4: Quality of Reporting for Three of the Most Commonly Used Techniques

In this section we describe the results of analysing the reporting of three of ICER's most commonly used tests: chi-squared, Mann-Whitney-Wilcoxon, and the t-test.

4.8.1 Reporting Statistical Tests. Table 7 shows the results of evaluating the descriptive statistics and test results using the criteria described in Section 3. Overall, the results are not encouraging, with very few papers reporting enough information to adequately describe the data and tests conducted.

A formal hypothesis was rarely stated clearly, although sometimes it could be inferred from the text. It is a concern that we found some cases where a test was reported with no explanation of its purpose or the data that it was applied to.

When describing the data set used in an analysis most papers included the sample size but fewer mentioned the size of each cell. Other descriptive statistics also tended to be inadequately reported. For example, measures of central tendency (mean or median) were reported in just over half the cases, and the measure of dispersion for t-tests (standard deviation) was reported in just over a quarter of the cases.

The reporting of the statistical analysis itself was often inadequate. The name of the test was always stated precisely for Mann-Whitney-Wilcoxon (except for one paper that named 'Wilcoxon' without stating whether it used the rank sum or signed rank version). On the other hand, for t-test and chi-squared, which are both families of tests, the precise name of the test was often not provided. Sometimes the particular test used could be inferred from the data but not naming the test reduces the readability of the reporting and the level of confidence in the analysis. In some cases there were vague hints of a test with statements such as 'a correlation was found' or 'there was no significant difference between groups', but no further information was provided. In the worst cases we

Table 7: Percentages of cases where each quality indicator was reported for t-tests, Mann Whitney U & Wilcoxon rank-sum & Wilcoxon signed-rank tests (MWW), and chi-squared tests (n = 32, 31, and 22, respectively)

Quality indicator	t-test (%)	MWW (%)	chi-sq (%)
hypothesis precisely stated	6	23	14
The data analysed			
sample size	94	93	91
size of cells/groups	72	83	59
mean	59	n/a	n/a
standard deviation	28	n/a	n/a
median	n/a	60	n/a
The statistical analysis			
precise name of test	59	97	9
one- or two-tailed	16	13	n/a
alpha-value	50	54	45
statistic	56	66	82
p-value	59	51	48
degrees of freedom	44	37	91
appropriate level of precision	59	60	82
statistical package	13	13	9

encountered a phenomenon that we have called 'alphabet soup': a list of symbols and values, with no mention of the test that was used.

Where tests could be either one or two-tailed, this was rarely specified. *Alpha*-values were not always provided, and it was even rarer to find them stated clearly before the analysis. For an example of good reporting in this regard, see Simon et al. [27]. Contrary to the recommendation of the APA Manual [2], papers often failed to report the exact *p*-value for significant results. The degrees of freedom were usually reported for chi-squared tests but much less frequently for t-tests and Mann-Whitney-Wilcoxon tests.

Further, even when the desired information was provided, the reader often had to hunt throughout the paper to find it, or even to infer it from the text (e.g., by computing the degrees of freedom from the number of rows and columns in a contingency table).

There are recognised guidelines for the reporting of numerical data, such as those of the APA. For example, percentages should be presented as integers and other numbers using two decimal places (except for *p*-values less than 0.01, which can use three or more decimal places). We often found results presented with inappropriate or unnecessary levels of precision; sometimes four or more decimal places were used for a statistical value.

In some disciplines it is customary to report the software that was used to run the statistical tests, as this provides further assistance to those who would like to check the results or conduct comparable tests on their own data. In our analysis we found that the statistical software used was rarely reported.

4.8.2 Using Randolph's Metric. Next, we applied Randolph's metric [18] to the papers reporting on these three techniques. Because almost all the papers in our data set were published after Randolph's call for improved statistical reporting, and because our entire data set consists of ICER papers (Randolph drew data only from the first year of ICER papers, and only four of those were included in the random sample he used for analysis), we were curious to see whether, by his definition, more of the papers describing our three techniques reported their analysis 'adequately'.[1]

Table 8: Comparison of our findings with Randolph's

Population	Adequate reporting	Inadequate reporting
Parametric tests		
Randolph parametric [18]	15 (60%)	10 (40%)
ICER t-test papers	11 (37%)	19 (63%)
Non-parametric tests		
Randolph non-parametric [18]	8 (73%)	3 (27%)
ICER chi-squared papers	13 (59%)	9 (41%)
ICER MWW papers	16 (55%)	13 (45%)

Our results are summarised along with Randolph's in table 8. Contrary to expectations, we found no significant difference between our results and his. Our null hypothesis for all tests was that there was no difference, with $\alpha = .05$. A chi-squared test of independence found no significant difference between the adequacy of the reporting in Randolph's parametric papers and ICER's t-test papers. For the nonparametric results, a Fisher two-tailed exact test was used because some of the expected cell frequencies were too low for a chi-squared test. The Fisher test found no significant difference between ICER's chi-squared papers and Randolph's nonparametric papers.

These results should be read with caution, for several reasons. First, Randolph's analysis of inferential statistics is restricted to papers that involve human participant data, while ours is not. In addition, he did not name the specific inferential techniques used in the papers he analysed, so we cannot undertake a technique-to-technique comparison. Finally, the sample sizes in these cases are relatively small. Still, the comparison does not indicate any improvement in the reporting of these statistics.

4.9 Other issues in the reporting of tests

In our analysis of the papers in our data set we identified a number of other issues with the use and reporting of statistical tests.

4.9.1 Data Preparation. A number of papers in our data set reported analysis that was based on non-numeric data such as computer programs, interviews, video, and open-ended questions. To conduct the quantitative analysis it was necessary to transform the qualitative data into quantitative data. In many cases this preprocessing stage was explained inadequately or not at all. Hundhausen and Brown [10] avoid this pitfall by giving a clear and detailed description of how their qualitative data was processed into quantitative data.

4.9.2 Assumptions. Many statistical tests should be applied only when certain assumptions have been met; for example, the level of data is appropriate (nominal, ordinal, or interval), the underlying

[1] None of the four ICER papers Randolph analysed was contained in our sets of chi-squared, MWW, or t-test papers.

distribution is normal, or variances are equal (homoscedasticity). When using such tests, it is important that the assumptions are explicitly stated to assure readers that they have been met. Where tests have been applied to check on the assumptions, these tests should also be fully described. We found in our analysis that the level of data was sometimes mentioned and there was some evidence of checking for normality or homoscedasticity; however, other assumptions were rarely discussed.

4.9.3 Obscure Tests. Our analysis revealed some statistical tests rarely seen in the computing education literature. Any test not commonly found should be explained at least briefly, with a reference for readers who would like more information about it. There is clearly some subjectivity in judging the relative obscurity of a test in computing education research, but many researchers will have some notion of which tests are used often, and our findings in this paper can help to confirm that. Patitsas et al. [17] provide an example of the use and explanation of relatively obscure tests.

4.9.4 Corrections for Multiple Tests. If a test is applied many times to a data set, there is a chance that one or more of the applications will achieve significance purely by chance. Adjustments such as the Bonferroni correction and the Benjamini-Hochberg procedure are designed to mitigate against that chance. We found a number of cases where multiple tests had been applied, but only a few mentioned the issue and used a test correction.

4.9.5 Statistical Significance and Effect Size. In statistics the word 'significance' carries a particular connotation, as at times has a very specific meaning. In reporting quantitative results it is important that this word is used only when reporting the result of a significance test. As a further complication, statistical significance does not necessarily imply 'practical' significance, by which we mean that the result has useful implications for practice.

The degree of practical significance can be conveyed by reporting the effect size. Effect size is gaining increasing prominence in many communities. There are various methods for computing effect size, both unstandardised (e.g., the difference between group means or odds ratio) and standardised. We did not look for the use of unstandardised methods, but, as described in Section 4.9.5, we found that standardised effect sizes were rarely reported.

5 DISCUSSION

Statistical analysis is an important tool for the computing education researcher. However, it can be a dangerous tool unless used properly. We found many issues with the use and reporting of statistics, some very concerning, and some aspects of our own prior papers that should have been better presented.

Why is this so? Is it lack of awareness? Is it lack of knowledge? Statistics is a broad and difficult area, with a huge learning curve for beginners. It is also possible that many reviewers lack either the background or the confidence to review papers with involved statistical analysis.

What can be done about these issues? Randolph attempted to raise awareness of the problem more than ten years ago; we are following in his footsteps. There are certainly examples of good practice as well as issues: notably Hanks [9], for its careful analysis and thoughtful considerations of its scope and limitations, and Toma and Vahrenhold [29], for its exemplary reporting in general. And for reviewers, it might make a difference simply to tell them that they are entitled to a clear hypothesis, a name for any test being used, mention of the descriptive statistics, and so forth, all organised and easy to locate within the paper.

6 LIMITATIONS

We acknowledge several limitations to this work. We did not include papers from other computing education venues such as SIGCSE, ITiCSE, LaTiCE, ACE, and Koli. We also did not include papers from journals such as ACM Transactions on Computing Education or Computer Science Education. The use of statistics in ICER papers may not be typical of these or other venues. Still, given the important role that ICER plays in empirical computing education research, we believe that an analysis of ICER's practice is of value to the community as a whole.

Further, much of our analysis focuses on three statistical techniques commonly used in ICER papers. It is possible that the usage and reporting of these techniques differ from those of other techniques, either at ICER or in general. But this is just a starting point. We encourage others to pick up this investigation and broaden and deepen it.

7 CONCLUSIONS AND FUTURE WORK

In this paper we report on a methodological review of all 270 ICER papers to date (2005-2018), in which we investigate the use and reporting of inferential statistics. Inferential statistics, an important aspect of the ICER community's research, are found in more than half of ICER papers, encompassing a wide variety of techniques.

There are excellent examples of reporting this analysis, but our deeper analysis of a sample of ICER papers (those using chi-squared, Mann-Whitney-Wilcoxon, and t-tests) reveals many issues. A comparison with the work done by Randolph more than ten years ago suggests that the reporting of statistical work in computing education research has not improved since then.

This study lays the groundwork for a number of possible future projects. For example, a similar analysis using a random sample from a broader set of journals and conferences, and/or techniques such as correlation and regression that we did not cover in this paper; an investigation into the use and reporting of effect size; or an examination of substantive issues such as the choice of test.

We are also interested in outreach activities that might help to raise the standard of statistics reporting. It is important that we lift our game – out of respect for the discipline, rigour of the research, faithfulness of the results, and ultimately to report research that can effect positive change for our students.

ACKNOWLEDGMENTS

We acknowledge the work of Robert McCartney, who has been part of the project team since its inception, but was not in a position to contribute to this particular paper.

REFERENCES

[1] American Educational Research Association. 2006. *Standards for Reporting on Empirical Social Science Research in AERA Publications.* American Educational Research Association. http://www.aera.net/Portals/38/docs/12ERv35n6_Standard4Report.pdf, (accessed 1 April 1 2019).

[2] American Psychological Association. 2009. *Publication Manual* (6th ed.). American Psychological Association, Washington, DC.

[3] Mousumi Banerjee, Michelle Capozzoli, Laura McSweeney, and Debajyoti Sinha. 1999. Beyond kappa: a review of interrater agreement measures. *Canadian Journal of Statistics* 27, 1 (1999), 3–23. https://doi.org/10.2307/3315487

[4] Lisa Bender, Gursimran Walia, Krishna Kambhampaty, Kendall E Nygard, and Travis E Nygard. 2012. Social sensitivity correlations with the effectiveness of team process performance: an empirical study. In *[Eighth] International Computing Education Research Conference (ICER 2012)*. ACM, New York, NY, USA, 39–46. http://doi.acm.org/10.1145/2361276.2361285

[5] Harris M Cooper. 1988. Organizing knowledge syntheses: a taxonomy of literature reviews. *Knowledge in Society* 1, 1 (1988), 104.

[6] Stephen Cooper, Karen Wang, Maya Israni, and Sheryl Sorby. 2015. Spatial skills training in introductory computing. In *11th International Computing Education Research Conference (ICER 2015)*. ACM, 13–20. https://doi.org/10.1145/2787622.2787728

[7] Mark Davies and Joseph L Fleiss. 1982. Measuring agreement for multinomial data. *Biometrics* 38, 4 (1982), 1047–1051. http://www.jstor.org/stable/2529886

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

[9] Brian Hanks. 2007. Problems encountered by novice pair programmers. In *Third International Computing Education Research Workshop (ICER 2007)*. 159–164. https://doi.org/10.1145/1288580.1288601

[10] Christopher D Hundhausen and Jonathan Lee Brown. 2005. Personalizing and discussing algorithms within CS1 studio experiences: an observational study. In *First International Computing Education Research Workshop (ICER 2005)*. 45–56. http://doi.acm.org/10.1145/1089786.1089791

[11] Matthew C Jadud and Brian Dorn. 2015. Aggregate compilation behavior: findings and implications from 27,698 users. In *11th International Computing Education Research Conference (ICER 2015)*. 131–139. http://doi.acm.org/10.1145/2787622.2787718

[12] Ximena López, Jorge Valenzuela, Miguel Nussbaum, and Chin-Chung Tsai. 2015. Some recommendations for the reporting of quantitative studies. *Computers & Education* 91 (2015), 106–110. https://doi.org/10.1016/j.compedu.2015.09.010

[13] Lauri Malmi, Judy Sheard, Simon, Roman Bednarik, Juha Helminen, Ari Korhonen, Niko Myller, Juha Sorva, and Ahmad Taherkhani. 2010. Characterizing research in computing education: a preliminary analysis of the literature. In *Sixth International Computing Education Research Workshop (ICER 2010)*. 3–12. http://doi.acm.org/10.1145/1839594.1839597

[14] Joe Miró Julià, David López, and Ricardo Alberich. 2012. Education and research: evidence of a dual life. In *[Eighth] International Computing Education Research Conference (ICER 2012)*. 17–22. http://doi.acm.org/10.1145/2361276.2361281

[15] Briana B Morrison, Brian Dorn, and Mark Guzdial. 2014. Measuring cognitive load in introductory CS: adaptation of an instrument. In *Tenth International Computing Education Research Conference (ICER 2014)*. 131–138. http://doi.acm.org/10.1145/2632320.2632348

[16] Kimberly A Neuendorf. 2016. *The Content Analysis Guidebook*. Sage.

[17] Elizabeth Patitsas, Jesse Berlin, Michelle Craig, and Steve Easterbrook. 2016. Evidence that computer science grades are not bimodal. In *12th International Computing Education Research Conference (ICER 2016)*. 113–121. http://doi.acm.org/10.1145/2960310.2960312

[18] JJ Randolph. 2007. *Computer science education research at the crossroads. A methodological review of computer science education research: 2000–2005.* Ph.D. Dissertation. Utah State University. https://archive.org/details/randolph_dissertation (accessed 15 February 2019).

[19] JJ Randolph. 2007. Findings from "A methodological review of the computer science education research: 2000–2005". *SIGCSE Bulletin* 39, 4 (Dec. 2007), 130. https://doi.org/10.1145/1345375.1345434

[20] JJ Randolph. 2009. A guide to writing the dissertation literature review. *Practical Assessment, Research & Evaluation* 14, 13 (2009), 1–13.

[21] JJ Randolph, R Bednarik, and N Myller. 2005. A methodological review of the articles published in the proceedings of Koli Calling 2001–2004. In *Koli Calling 2005 Conference on Computer Science Education (Koli Calling 2005)*. 103–109. Linked from https://www.kolicalling.fi/index.php/previous-koli-calling-conferences (Retrieved 15 February 2019).

[22] JJ Randolph, G Julnes, E Sutinen, and S Lehman. 2008. A methodological review of computer science education research. *Journal of Information Technology Education* 7 (2008), 135–162.

[23] Judy Sheard, Simon, Margaret Hamilton, and Jan Lönnberg. 2009. Analysis of research into the teaching and learning of programming. In *Fifth International Computing Education Research Workshop (ICER 2009)*. 93–104. http://doi.acm.org/10.1145/1584322.1584334

[24] Simon. 2007. A classification of recent Australasian computing education publications. *Computer Science Education* 17, 3 (2007), 155–169.

[25] Simon. 2007. Koli Calling comes of age: an analysis. In *Seventh Baltic Sea Conference on Computing Education Research (Koli Calling 2007)*. 119–126. http://dl.acm.org/citation.cfm?id=2449323.2449336

[26] Simon, Judy Sheard, Angela Carbone, Michael de Raadt, Margaret Hamilton, Raymond Lister, and Errol Thompson. 2008. Eight years of computing education papers at NACCQ. In *21st Annual Conference of the National Advisory Committee on Computing Qualifications (NACCQ 2008)*. National Advisory Committee on Computing Qualifications, 101–107. https://www.citrenz.ac.nz/conferences/2008/101.pdf

[27] Beth Simon, Tzu-Yi Chen, Gary Lewandowski, Robert McCartney, and Kate Sanders. 2006. Commonsense computing: what students know before we teach (episode 1: sorting). In *Second International Computing Education Research Workshop (ICER 2006)*. 29–40. http://doi.acm.org/10.1145/1151588.1151594

[28] Teemu Sirkiä and Juha Sorva. 2015. How do students use program visualizations within an interactive ebook?. In *11th International Computing Education Research Conference (ICER 2015)*. 179–188. http://doi.acm.org/10.1145/2787622.2787719

[29] Laura Toma and Jan Vahrenhold. 2018. Self-efficacy, cognitive load, and emotional reactions in collaborative algorithms labs – a case study. In *14th International Computing Education Research Conference (ICER 2018)*. 1–10. http://doi.acm.org/10.1145/3230977.3230980

[30] David W Valentine. 2004. CS educational research: a meta-analysis of SIGCSE Technical Symposium proceedings. In *35th ACM Technical Symposium on Computer Science Education (SIGCSE 2004)*. ACM, 255–259.

[31] Jeffrey C Valentine and Harris Cooper. 2003. Effect size substantive interpretation guidelines: Issues in the interpretation of effect sizes. *Washington, DC: What Works Clearinghouse* (2003), 1–7.