Computers & Education 91 (2015) 106–110

Contents lists available at ScienceDirect

Computers & Education

journal homepage: www.elsevier.com/locate/compedu

Editorial

Some recommendations for the reporting of quantitative studies

The quality of research design and results reporting are of paramount importance when judging and ultimately accepting academic articles for publication. The pertinence of methodological procedures and the appropriateness of statistical analysis are not only relevant to defend the robustness of the research results being presented; more importantly, they are relevant to understand the contribution of these results to the broader field of study.

The academic debate has become increasingly controversial regarding the use of widely accepted statistical procedures, such as null hypothesis significance testing and related techniques (Sharpe, 2013; Trafimow & Marks, 2015). Quantitative experts have criticised the over-reliance on these procedures as they not only provide insufficient quantitative information on the data; more worryingly, they may also mislead the interpretation of results by fellow researchers (Kline, 2013). The debate has also been extended to other procedures, suggesting that misuse of appropriate statistical practices may stem from insufficient knowledge of further statistical concepts (such as statistical power and test assumptions) or from a lack of awareness of more modern and advanced statistical techniques (Sharpe, 2013). In other cases, such misuse may be associated with theoretical and/or operational difficulties when planning or carrying out the study (Cumming, 2014).

This editorial aims to provide guidelines to encourage a clearer and more complete reporting of research outputs for quantitative studies. Research outputs, however, are directly determined by the research design and data analysis. As Pierson (2004) states, "*No amount of rewriting, creative data presentation, or statistical manipulation can make up for the fact that the study used the wrong model or study design, collected data in a manner that would not allow a meaningful examination of the hypothesis, or made too few measurements to permit confident conclusions to be drawn*" (p. 1250). Hence, these guidelines are also an invitation to reflect on the suitability of research design decisions and the suitability of data analysis procedures for the research aims within a quantitative approach.

It is not our intent to suggest that we will accept manuscripts using a quantitative method over a qualitative one. We deeply value a plurality of epistemological approaches and research designs that suit the vast variety of questions that are intrinsic to the complex nature of our field of study. However, given the large amount of research studies using a quantitative design submitted to Computers & Education, it is important to make explicit some specific guidelines that will contribute to the quality of such manuscripts.

We will, in due course, publish an editorial focussed on enhancing the quality of submissions that are based on research that adopts a qualitative approach.

As a general recommendation, manuscripts based on a quantitative approach that are submitted for publication, besides having an appropriate number of significant figures, should report detailed information to allow a knowledgeable reader to a) replicate the study; b) assess the rigorousness of the research design; and c) evaluate the robustness of the results and generalisability of the conclusions. In order to do so, authors should consider the following aspects:

1. Regarding participants

1.1. Using representative samples

A representative sample is a subset of elements, the characteristics of which reflect the whole population from which it has been drawn. Using a representative sample is the only way to generalise the results for the population as a whole. Randomisation is key to ensuring the representativeness of the sample. Randomisation in this sense means the random selection of subjects from a given sample or the random allocation of subjects to different experimental groups. This allows for an estimation of the error for the sample, which is necessary for determining the sample size.

http://dx.doi.org/10.1016/j.compedu.2015.09.010

It should be clear in the manuscript that the study is using a representative sample. It is important to describe the universe from which the population has been drawn (size of the population, main characteristics, etc.), as well as the sampling techniques that have been used (systematic, stratified, cluster, etc.). Different sampling techniques require different variance formulae in order to derive the confidence intervals. If the conclusions of the study involve generalising for subgroups, then the sample size should be representative at the subgroup level. It is also important to indicate the confidence level and margin of error (confidence interval) used to calculate the sample size (see Cohen, Manion, & Morrison, 2011; Kerlinger & Lee, 2000).

1.2. Not using representative samples

In education it is often not possible to obtain a representative sample. In these cases, the manuscripts must not only indicate the total number of participants and their characteristics; they must also justify and specify the inclusion and exclusion criteria for the sample, as well as justifying the size of the sample (i.e. why that number of participants and not another). Furthermore, it is important to report the Effect Size (Sullivan & Feinn, 2012; Trusty, Thompson, & Petrocelli, 2004), which can be detected using a given Power and the study's sample size (Faul, Erdfelder, Lang, & Buchner, 2013) (see Section 3.1). Using this information, the design sensitivity can be assessed in order to observe the phenomenon in question (see Cohen, 1992; Ellis, 2010).

2. Regarding instruments

One criterion for assessing the robustness of the study is to provide evidence to show that the instrument used to measure the variables is valid and reliable, i.e. it measures what it is supposed to measure and does so with a suitable degree of accuracy (DeVellis, 2012). Given that these properties stem from the specific use of the instrument in a set population, regardless of whether the instrument has already been validated or has been created within the framework of the study, it is important to at least report the estimate.

As well as content validity (whether or not the test covers a representative sample of the variable to be measured) and predictive validity (the degree to which the variable of interest can be effectively predicted), construct validity is a key element in the robustness of the instrument when measuring variables that are not directly observed as it reveals that the instrument measures the construct in question. Because of this, whenever such variables are measured it is essential to reference the validation study. Should such a study not exist, the factorial structure may be reported. If this is done using Exploratory Factor Analysis (EFA), the main parameters must be reported: number of factors extracted, total variance explained by extracted factors, Kaiser Meyer Olkin index (KMO), sample size, and ratio of number of participants to number of variables factored (Henson & Roberts, 2006). Should confirmatory Factor Analysis (CFA) be used, it is important to explicitly indicate whether the results correspond to the validation study or whether they refer to data obtained from the study that is reported. In the latter case, both the model that is tested as well as the relevant fit indices must be reported (see Schreiber, Nora, Stage, Barlow, & King, 2006).

However, it is not enough for the instrument to be valid; it must also be reliable, i.e. it must accurately measure what it aims to measure. While there are several alternatives for demonstrating reliability, such as the *composite reliability value* (usually calculated in conjunction with structural equation modelling) (Raykov, 1997), the index that is most widely used by researchers continues to be Cronbach's alpha, despite being an internal consistency estimate and not a direct measure of reliability (Henson, 2001). The main weakness of Cronbach's alpha is that it overestimates internal consistency due to the number of items (Osburn, 2000). Despite this limitation, the alternative techniques proposed for estimating reliability, such as the *composite reliability value*, do not manage to reveal consequential differences (Peterson & Kim, 2013).

Traditionally, an acceptable level for Cronbach's alpha is considered to be >0.7, although authors slightly modify this criterion depending on the type of research. For example, if the results of the research have an impact on important decisions that are to be made, a much higher level of reliability is required (see Cortina, 1993; George & Mallery, 2003; Kerlinger & Lee, 2000; Streiner, 2003). Despite this, and further to the above, it is also important to bear in mind two additional points when calculating and reporting Cronbach's alpha. The first of these points is that Cronbach's alpha supposes the one-dimensional nature of the construct that is being evaluated, i.e. that the set of items or indicators only measure a single dimension (Barbaranelli, Lee, Vellone, & Riegel, 2015). Given this, if the construct that is evaluated has more than one dimension (e.g. subscales), the reliability index for each of these dimensions must be indicated. The second point that must be borne in mind is that this index is especially sensitive to the number of items included in a scale, meaning that an increase in the number of items may lead to an increase in Cronbach's alpha.

3. Regarding data collection and analysis

Statistical analyses depend on the research design and type of data that are collected. It is therefore important to choose the statistical test that best fits the nature of the data. For example, if a categorical variable is to be predicted based on a series of predictor variables, a logistic regression must be used instead of a linear regression. In this sense, the manuscript should provide clear evidence that the data analysis methods used are sound and coherent with the study's objectives. Methods should be described and justified, indicating that tests assumptions have been verified. There are several statistical manuals (e.g. Cohen et al., 2011; Kerlinger & Lee, 2000) and online resources (e.g. http://bama.ua.edu/~jleeper/627/choosestat.html)

that aid the decision making process in order to choose the appropriate analysis according to the type of variables being measured and specific characteristics of the data (e.g. unequal sample sizes and violation of assumptions).

3.1. Checking on assumptions

It should also be borne in mind that different statistical tests have different assumptions that need to be met in order to return valid results. Parametric tests such as the t-test, ANOVA and Pearson's correlation are based on four assumptions: a) Normal distribution of data (more specifically the normal distribution of the residuals); b) Homogeneity of variance; c) Intervalar data; and d) Independence of data (or error, in the case of regressions) (Osborne & Waters, 2002; Williams, Gómez, & Kurkiewicz, 2013). When parametric assumptions are violated, several options are available:

1. Using a test that corrects for a particular assumption violation, for example, the Welch's t-test (Welch, 1947) when variances of two groups are not equal; or Games–Howell post hoc test when there is no homogeneity of variance in an ANOVA (Games & Howell, 1976);
2. Using non-parametric tests, which converts interval data into rank-ordered data (for example, using the Whitney-Mann-Wilcoxon test (Fay & Proschan, 2010) instead of the t-test when independent data are not normally distributed; using the Kruskal–Wallis test (Spurrier, 2003) instead of the ANOVA when there are more than two groups);
3. Performing data corrections, for example, removing outliers or applying mathematical transformations (Cousineau & Chartier, 2010; Leys, Ley, Klein, Bernard, & Licata, 2013; Rousseeuw & Leroy, 2005).

In these cases, such procedures must be clearly described and justified in order to enable the reader to assess the outreach and limitations of the study's results.

3.2. Defining dependent and independent measurements

Data analysis procedures are determined by the dependence or independence of the data measurements. Data is independent if measurements from one group are unrelated to measurements from the other group. Similarly, measurements are dependent when measurements from one group are paired with the measurements from the other group, for example, when doing a pre-post test, or when studying the effects of different treatments using the same subjects (see Cohen et al., 2011). Independent designs are frequently used, as using a dependent design can be expensive and/or difficult to carry out. However, when using a dependent design, the random variance that is due to individual differences is strongly reduced, increasing the power of the study (Kirk, 2009) (see Section 4.1). Different statistical tests should be used to determine the difference between groups, according to the type of design. For example, comparing two means in a dependent design requires a paired t-test; if measurements are independent, an independent t-test can be used (Cohen et al., 2011). Invalid conclusions may be drawn if data are analysed while dismissing the type of design used (Hailman & Strier, 2006).

4. Regarding results

4.1. Hypothesis testing: reporting p-values, effect sizes and power

Despite being criticised for its limited value for conveying relevant quantitative information, Null Hypothesis Statistical Testing (NHST) is widely used among researchers (Sharpe, 2013). A null hypothesis usually takes forms such as “no effect”, “no difference” or “no correlation” in a population. An alternative hypothesis is the denial of the null hypothesis and usually (but not always) corresponds to what the study expects to demonstrate (i.e. the existence of an effect, a difference or a correlation). Good hypotheses are testable; they should be stated clearly and specifically and should be aligned with the study's aims (Hailman & Strier, 2006).

When testing a null hypothesis, a p-value is calculated to determine the significance level, i.e. the probability of rejecting a null hypothesis that is true. Researchers traditionally establish 5% as the significance level, thus any value smaller than that is considered statistically significant (Cohen et al., 2011). There has been some confusion in interpreting the p-value, as some consider statistical significance as a measure of a substantive significance. It is important to stress that if a null hypothesis is rejected, a significant p-value only indicates that there is a low probability of a “no effect” situation, and it does not provide any evidence about the theoretical meaningfulness or magnitude of the effect (Sharpe, 2013).

When reporting p-values, Effect Size estimates (Fritz, Morris, & Richler, 2012) and the power (Ellis, 2010) of the study should also be communicated. Effect size groups a number of indices that provide a quantitative description of the magnitude of the observed effects. Effect sizes are independent of sample size, unlike significance tests. For example, studies with the same mean and standard deviation will have the same effect size estimates but different statistical significance depending on the size of the sample. Different estimates of effect size are frequently associated with the type of statistical analysis performed. Cohen's *d* is most commonly found when conducting t-tests, partial eta squared has mostly been reported in studies using ANOVA, while adjusted R^2 has been associated with regression analyses (Fritz et al., 2012).

For each reported effect size, it is recommended that a confidence interval (CI) is presented in order for the reader to assess the accuracy of the estimation of the effect size (American Psychological Association, 2010). However, calculating these CIs is not simple as most measures of effect size are not centrally distributed (Fritz et al., 2012). This means that the distribution of the plausible values of effect size in the population is not symmetrically distributed around 0, which impacts the way in which the lower and upper limits of the CI are calculated. Online sites such as www.thenewstatistics.com offer free downloadable resources that calculate CI.

Effect size estimates should not only be communicated but also interpreted in light of the theoretical or practical framework of the study. A small effect may be very relevant in contexts in which the dependent variable is very resistant to change (Fritz et al., 2012). Moreover, effect sizes are a valuable input for meta-analysis studies as they enable comparisons of effects across studies using different samples and data analysis techniques (Ellis, 2010).

Reporting effect sizes also communicates information about the power of the study. Power refers to the sensitivity of the study design to detect an effect and corresponds to the probability of correctly rejecting a null hypothesis when the alternative hypothesis is true (Ellis, 2010). A good practice is to define the power of the study a priori to collecting the data in order to calculate the sample size needed to detect a determined effect size (Cumming, 2014). It is recommended reporting and discussing post hoc statistical power when the final sample size and/or the effect size obtained do not correspond to those expected when designing the study. Studies may find large effects that are non-significant, which suggests the need for further research with greater power. Likewise, using large samples may lead to significant effects that are small. In this case, the observed effects may be overvalued (Fritz et al., 2012).

Finally, it is important to bear in mind that preference should be given to indices that provide a more robust result. In the case of effect size for multiple regressions, for example, it is suggested that authors take the value of the adjusted R^2 as an indicator of the model's suitability. This is because the adjusted R^2 takes into consideration both the sample size as well as the number of parameters included in the model. It is also suggested that authors report the β value (see Draper & Smith, 2014).

4.2. Using Bayesian procedures

The use of modern techniques for data analysis has been proposed to overcome problems of NHST, such as Bayesian methods. Bayesian methods provide many advantages, such as being able to include previous knowledge about parameters (known as *prior*), handle uncertainty with small data sets, and use unbalanced samples with missing data (Wasserman, 2004). Due to its slow adoption (Sharpe, 2013), it is advisable to explain as clearly as possible the rationale of the analysis and the advantages of the method used when using Bayesian methods. It is also advisable to describe the model and its parameters, the likelihood function, the prior, and the chain convergence (Kruschke, 2013). This will allow fellow researchers to contextualise the findings in light of the priors used, replicate results and potentially update priors for future research.

5. Summary

This set of suggestions aim to provide examples of a series of elements that may significantly contribute towards demonstrating the robustness of quantitative results. It is therefore not a methodological guide but instead a guide that acts as a reminder of some basic principles when reporting this type of study. The research in Computers & Education has moved beyond simple survey data and any research paper sent for review must triangulate the results from a variety of sources. We believe it is essential to highlight that a high quality paper with a quantitative approach must provide evidence not only of the robustness of its theoretical framework, but also of its design and the statistical procedures chosen in order to answer the research question. In this sense, the statistical report must satisfy the following general principle: demonstrate the solidity of the treatment of the data so as to allow the robustness of the results and generalisability of conclusions to be evaluated.

References

American Psychological Association. (2010). *The publication manual of the American psychological association* (6th ed.). Washington, DC: American Psychological Association.

Barbaranelli, C., Lee, C. S., Vellone, E., & Riegel, B. (2015). The Problem With Cronbach's alpha: comment on Sijtsma and van der Ark. *Nursing Research, 64*(2), 140–145.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*(1), 155–159.

Cohen, L., Manion, L., & Morrison, K. (2011). *Research methods in education* (7th ed.). Oxon, [England]: Routledge. Milton Park, Abingdon.

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology, 78*, 98–104.

Cousineau, D., & Chartier, S. (2010). Outliers detection and treatment: a review. *International Journal of Psychological Research, 3*(1), 58–67.

Cumming, G. (2014). The new statistics: why and how. *Psychological Science, 25*(1), 7–29.

DeVellis, R. F. (2012). *Scale development: Theory and applications*. Los Angeles, CA: Sage Publications.

Draper, N. R., & Smith, H. (2014). *Applied regression analysis* (3rd ed.). New York: John Wiley & Sons.

Ellis, P. D. (2010). *The essential guide to effect sizes: Statistical power, meta-analysis, and the interpretation of research results*. Cambridge: Cambridge University Press.

Faul, F., Erdfelder, E., Lang, A., & Buchner, A. (2013). G* power 3.1. 7: a flexible statistical power analysis program for the social, behavioral and biomedical sciences. *Behavior Research Methods, 39*, 175–191.

Fay, M. P., & Proschan, M. A. (2010). Wilcoxon–Mann–Whitney or t-test? On assumptions for hypothesis tests and multiple interpretations of decision rules. *Statistics Surveys, 4*, 1–39. http://dx.doi.org/10.1214/09-SS051.

Fritz, C. O., Morris, P. E., & Richler, J. J. (2012). Effect size estimates: current use, calculations and interpretation. *Journal of Experimental Psychology: General, 141*, 2–18.

Games, P., & Howell, J. F. (1976). Pairwise multiple comparison procedures with unequal N's and/or variances: a Monte Carlo study. *Journal of Educational and Behavioral Statistics, 1*(2), 113–125.
George, D., & Mallery, M. (2003). *Using SPSS for Windows step by step: A simple guide and reference*. Boston, MA: Allyn & Bacon.
Hailman, J. P., & Strier, K. B. (2006). *Planning, proposing, and presenting science effectively* (2nd ed.). Cambridge: Cambridge University Press.
Henson, R. K. (2001). Understanding internal consistency reliability estimates: a conceptual primer on coefficient alpha. *Measurement and Evaluation in Counseling and Development, 34*(3), 177–189.
Henson, R. K., & Roberts, J. K. (2006). Use of exploratory factor analysis in published research common errors and some comment on improved practice. *Educational and Psychological Measurement, 66*(3), 393–416.
Kerlinger, F. N., & Lee, H. B. (2000). *Foundations of behavioral research* (4th ed.). Holt, NY: Harcourt College Publishers.
Kirk, R. E. (2009). Experimental design. In R. Millsap, & A. Maydeu-Olivares (Eds.), *Sage handbook of quantitative methods in psychology* (pp. 23–45). Thousand Oaks, CA: Sage.
Kline, R. B. (2013). *Beyond significance testing: Statistics reform in the behavioral sciences* (2nd ed.). Washington, DC: American Psychological Association. http://dx.doi.org/10.1037/14136-000.
Kruschke, J. K. (2013). Bayesian estimation supersedes the t test. *Journal of Experimental Psychology: General, 142*(2), 573–603. http://dx.doi.org/10.1037/a0029146.
Leys, C., Ley, C., Klein, O., Bernard, P., & Licata, L. (2013). Detecting outliers: do not use standard deviation around the mean, use absolute deviation around the median. *Journal of Experimental Social Psychology, 49*(4), 764–766. http://dx.doi.org/10.1016/j.jesp.2013.03.013.
Osborne, J., & Waters, E. (2002). Four assumptions of multiple regression that researchers should always test. *Practical Assessment, Research & Evaluation, 8*(2). Retrieved May 30, 2015 from http://PAREonline.net/getvn.asp?v=8&n=2http://pareonline.net/getvn.asp?v=8&n=2.
Osburn, H. G. (2000). Coefficient alpha and related internal consistency reliability coefficients. *Psychological Methods, 5*(3), 343–355. http://dx.doi.org/10.1037/1082-989X.5.3.343.
Peterson, R. A., & Kim, Y. (2013). On the relationship between coefficient alpha and composite reliability. *Journal of Applied Psychology, 98*(1), 194–198. http://dx.doi.org/10.1037/a0030767.
Pierson, D. (2004). The top 10 reasons why manuscripts are not accepted for publication. *Respiratory Care, 49*, 1246–1252.
Raykov, T. (1997). Estimation of composite reliability for congeneric measures. *Applied Psychological Measurement, 21*(2), 173–184. http://dx.doi.org/10.1177/01466216970212006.
Rousseeuw, P. J., & Leroy, A. M. (2005). *Robust regression and outlier detection* (Vol. 589). John Wiley & Sons.
Schreiber, J. B., Nora, A., Stage, F. K., Barlow, E. A., & King, J. (2006). Reporting structural equation modeling and confirmatory factor analysis results: a review. *The Journal of Educational Research, 99*(6), 323–338.
Sharpe, D. (2013). Why the resistance to statistical innovations? Bridging the communication gap. *Psychological Methods, 18*(4), 572–582.
Spurrier, J. D. (2003). On the null distribution of the Kruskal–Wallis statistic. *Journal of Nonparametric Statistics, 15*(6), 685–691. http://dx.doi.org/10.1080/10485250310001634719.
Streiner, D. L. (2003). Starting at the beginning: an introduction to coefficient alpha and internal consistency. *Journal of Personality Assessment, 80*(1), 99–103.
Sullivan, G. M., & Feinn, R. (2012). Using effect size—or why the *P* value is not enough. *Journal of Graduate Medical Education, 4*(3), 279–282. http://dx.doi.org/10.4300/JGME-D-12-00156.1.
Trafimow, D., & Marks, M. (2015). Editorial. *Basic and Applied Social Psychology, 37*(1), 1–2.
Trusty, J., Thompson, B., & Petrocelli, J. V. (2004). Practical guide for reporting effect size in quantitative research in the journal of counseling & development. *Journal of Counseling & Development, 82*(1), 107–110. http://dx.doi.org/10.1002/j.1556-6678.2004.tb00291.x.
Wasserman, L. (2004). *All of statistics: A concise course in statistical inference*. New York: Springer-Verlag.
Welch, B. L. (1947). The generalization of "student's" problem when several different population variances are involved. *Biometrika, 34*(1–2), 28–35. http://dx.doi.org/10.1093/biomet/34.1-2.28.
Williams, M. N., Gómez, C. A., & Kurkiewicz, D. (2013). Assumptions of multiple regression: correcting two misconceptions. *Practical Assessment, Research & Evaluation, 18*(11), 1–14.

Ximena López
Universidad Roma III, Rome, Italy

Jorge Valenzuela
Centro de Estudios Avanzados, Universidad de Playa Ancha, Chile

Miguel Nussbaum *
Pontificia Universidad Católica de Chile, Chile

Chin-Chung Tsai
National Taiwan University of Science and Technology, Taiwan
* Corresponding author.
E-mail address: mn@ing.puc.cl

Available online 25 September 2015