Short Report

Quantitative Methods: Mistakes to Avoid

Christopher Pirotto
Fukui University of Technology

Robert Dykes
Jin-ai University

The implementation of standards for reporting quantitative methods results in many well-known journals has created the image that language learning researchers are correctly reporting quantitative research. However, there are still strong suggestions being made that better methodological training for researchers within the language learning field is necessary. Drawing on the authors' personal observations of conference presentations as well as conversations they had with many novice researchers, this paper outlines five common mistakes that researchers make in our field when using quantitative research methods. These mistakes are as follows: assuming non-statistically significant results are irrelevant, overvaluing statistically significant results and not reporting effect size, using incorrect effect size scales, forgetting to report certain statistical values, and changing the nature of a variable to fit a certain statistical test. This paper discusses each of these mistakes, provides reasons why these are mistakes, and includes advice on how to correct these mistakes. Reducing the number of times these mistakes are made will not only help to strengthen the quality of quantitative research within our field, but will also allow us to have more confidence in the decisions we make involving our classrooms and language learners.

Applied linguistic and TESOL research often informs teaching methodologies, materials development, and the decisions instructors make in the classroom. This research then has a direct impact on the language learning potential of language learners. Therefore, it is imperative that research be conducted in the most rigorous way possible. Byrnes (2013) stated that within the applied linguistics field, the level of methodological awareness has been increasing and has taken a "methodological turn" for the better. There is plenty of evidence to support Byrnes' claim including the calling for and implementation of standards for reporting quantitative methods and results in language learning research (Norris, Plonsky, Ross, & Schoonen, 2015), recommendations for improving the statistical knowledge of graduate students (Gonulal, Loewen, & Plonsky, 2017), and the introduction of new statistical techniques such as robust statistics (Larson-Hall & Herrington, 2010) and Bayesian statistics (Mackey & Ross, 2015). However, the evidence providing proof of the methodological turn is the same evidence that the language learning research field still has much to learn about quantitative methods.

This paper stems from quantitative mistakes the authors made in their own research, and it is possible that other researchers have made similar mistakes. The authors provide reasons as to why these are mistakes and also provide solutions, alternatives, or advice on how to deal with them. A glossary of statistical terminology is included at the end of this article (Appendix), and terms in it are italicized in the paper for easy recognition.

Mistake #1 – Overvaluing Statistical Significance

A novice researcher might believe that results which are not statistically significant, or even results with a low *effect size* are worthless and should not be published. While there is a bias towards publishing research that includes statistically significant findings (Rothstein, Sutton, & Borenstein, 2005), it still is possible to publish with results lacking statistical significance. In fact, there is tremendous value in knowing that a certain treatment or methodology has little or no effect on language learning. It is arguably just as important to know what does not work as what does.

Conversely, obtaining statistically significant results in and of itself does not justify the need for pedagogical changes. In fact, numerous researchers in our field have been calling for a shift away from *null hypothesis significance testing* and the overvaluing of statistical significance and towards the reporting and use of effect sizes (Larson-Hall & Plonsky, 2015; Norris, 2015; Plonsky, 2015b).

Effect size is not binary like statistical significance; it is a statistical measure that quantifies the size (or magnitude) of a given phenomenon e.g., indicating the difference between two groups or the strength of a *correlation* between two variables. In language learning, effect size can help show just how effective a given treatment or teaching method is for language learners. Calculating effect size is very simple, and a quick internet search for "effect size calculator" will yield several helpful tools.

As Table 1 shows, as a study's sample size increases the chance of obtaining statistical significance also increases, despite the *mean* and *standard deviations* not changing. However, the effect size value (*Cohen's* d) does not change. There is nothing wrong with reporting statistical significance, but effect size should also be reported and considered when drawing any conclusions.

Table 1

Effect of Sample Size on p *Value and Cohen's* d

Sample size of each group	Mean and standard deviation of group 1	Mean and standard deviation of group 2	p value	Cohen's d
10	46.4, 4.0	45.0, 3.2	0.40	0.39
50	46.4, 4.0	45.0, 3.2	0.06	0.39
200	46.4, 4.0	45.0, 3.2	0.01	0.39

Mistake #2 –Interpreting Effect Size

The two most common effect size variables are Cohen's *d*, which is used when comparing the means of two groups, and *Pearson's* r *correlation*, which is used to measure the strength of relationship between two variables. Calculating the effect size is rather easy with online tools, but interpreting them can be rather difficult. Is a Cohen's *d* value of 0.50 indicative of a small or large effect size? Table 2 shows two different sources that are cited for effect size interpretation, Cohen (1988) and Plonsky and Oswald (2014). The use of either of these scales is a step in the right direction; however, Cohen's (1988) scale was not created with applied linguistics or foreign language teaching in mind. In fact, Cohen (1988)

Table 2
Comparison of Effect Size Scales

Size	Cohen (1988)	Plonsky & Oswald (2014)
Small	$d = .2, r = .1$	$d = .45, r = .25$
Medium	$d = .5, r = .3$	$d = .71, r = .37$
Large	$d = .8, r = .5$	$d = 1.08, r = .54$

stated that effect sizes should be understood within the context of a specific field. This is exactly what Plonsky and Oswald (2014) did when they created their empirically based and field-specific effect size scale for applied linguistics. As an academic field, we should be moving away from Cohen's (1988) effect size scale and use Plonsky and Oswald's (2014) scale instead.

Mistake #3 – "Forgetting" to Report

"Forgetting" to report certain statistical items is a common mistake. Besides effect size, other unreported or otherwise unaddressed items include standard deviations (Plonsky, 2013), *confidence intervals* (Larson-Hall & Plonsky, 2015) and *assumptions* (Plonsky, 2015a). The reason for not reporting might be due to either not being aware of what should be reported or not understanding the significance of reporting a given value. Standard deviation tells us how far a group is dispersed from the mean, and this provides a picture of what is happening to the entire population. A low standard deviation value provides evidence that data points are closely grouped around the mean, whereas a high standard deviation value suggests that data points are spread out much further from the mean. This is important when doing research, because if a standard deviation increases by a large amount after a given treatment, it provides evidence that the treatment may not have been effective for a portion of the participants. Reporting standard deviation is also necessary for a study's data to be included in meta-analyses. Confidence intervals are important because participants of a study usually only represent a fraction of the entire population. Therefore, if the study was replicated with different participants, there is a high likelihood that the results would be different. A 95% confidence interval would show us a range of values that contain (with a 95% likelihood) the population's true value, whether that value be the mean, effect size, or some other statistical value. Each statistical test has its own assumptions that need to be met or addressed. If assumptions cannot be met, there are often alternative statistical tests that can be used. Addressing these assumptions in two or three lines of an article can give the readers confidence that the appropriate statistical test was run.

Mistake #4 – Variable Type

When deciding what kind of statistical test to run, the nature of the variables needs to be taken into account. Second language research often relies on a limited number of statistical procedures, mostly analysis of variance (*ANOVA*), correlations, and t-*tests* (Plonsky, 2013). As Plonsky (2015a) points out, "It is not uncommon to find researchers that convert intervally measured (independent) variables into categorical ones in order for the data to fit into an ANOVA model" (p. 3).

Another field-specific example would be grouping students into low-level and high-level proficiency groups when there is nothing different about the treatment or nature of these groups. A researcher might choose to do this because having two groups allows for the use of a *t* -test, one of the most common and easily interpretable statistical tests. In reality, these low-level and high-level groups are often defined by test scores. Therefore, instead of categorizing proficiency into a certain number of groups, it might be better to define proficiency by the test scores. Doing this will make it so that a *t*-test is no longer possible, but it provides for a more accurate representation of proficiency. It is important to define variables by their true nature and not mold them to fit a specific statistical test.

Final Words

Making mistakes and learning from them is one of the best forms of education. The purpose of this paper is to help improve the quality of quantitative research in our field. Improving how we report quantitative research findings will better inform our teaching methods, positively impact our classrooms, and in turn have a positive effect on the language learning process of our learners.

References

Byrnes, H. (2013). Notes from the editor. *The Modern Language Journal*, *97*(4), 825-827. https://doi.org/10.1111/j.1540-4781.2013.12051.x

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences.* Hillsdale, NJ: Lawrence Erlbaum.

Gonulal, T., Loewen, S., & Plonsky, L. (2017). The development of statistical literacy in applied linguistics graduate students. *International Journal of Applied Linguistics*, *168*(1), 4-32. https://doi.org/10.1075/itl.168.1.01gon

Larson-Hall, J. (2016). *A guide to doing statistics in second language research using SPSS and R*. New York, NY: Routledge.

Larson-Hall, J., & Herrington, R. (2010). Improving data analysis in second language acquisition by utilizing modern developments in applied statistics. *Applied Linguistics*, *31*(1), 368-390. https://doi.org/10.1093/applin/amp038

Larson-Hall, J., & Plonsky, L. (2015). Reporting and interpreting quantitative research findings: What gets reported and recommendations for the field. *Language Learning*, *65*(Supp. 1), 127-159. https://doi.org/10.1111/lang.12115

Mackey, B., & Ross, S. (2015). Bayesian informative hypothesis testing. In L. Plonsky (Ed.), *Advancing quantitative methods in second language research* (pp. 329-345). New York, NY: Routledge.

Norris, J. (2015). Statistical significance testing in second language research: Basic problems and suggestions for reform. *Language Learning*, *65*(Supp. 1), 97-126. https://doi.org/10.1111/lang.12114

Norris, J., Plonsky, L., Ross, S., & Schoonen, R. (2015). Guidelines for reporting quantitative methods and results in primary research. *Language Learning*, *65*(2), 470-476. https://doi.org/10.1111/lang.12104

Plonsky, L. (2013). Study quality in SLA: An assessment of designs, analyses, and reporting practices in quantitative L2 research. *Studies in Second Language Acquisition*, *35*(4), 655-687. https://doi.org/10.1017/S0272263113000399

Plonsky, L. (2014). Study quality in quantitative L2 research (1990-2010): A methodological synthesis and call for reform. *Modern Language Journal, 98*(1), 450-470. https://doi.org/10.1111/j.1540-4781.2014.12058.x

Plonsky, L. (2015a). Introduction. In L. Plonsky (Ed.), *Advancing quantitative methods in second language research* (pp. 3-8). New York, NY: Routledge.

Plonsky, L. (2015b). Statistical power, *p* values, descriptive statistics, and effect sizes: A "back-to-basics" approach to advancing quantitative methods in L2 research. In L. Plonsky (Ed.), *Advancing quantitative methods in second language research* (pp. 23-45). New York, NY: Routledge.

Plonsky, L., & Oswald, F. L. (2014). How big is "big"? Interpreting effect sizes in L2 research. *Language Learning, 64*(4), 878-912. https://doi.org/10.1111/lang.12079

Rothstein, H., Sutton, A., & Borenstein, M. (2005). *Publication bias in meta-analysis' Prevention, assessment, and adjustments.* West Sussex, England; John Wiley.

Author bios

Christopher Pirotto *is an assistant professor at Fukui University of Technology. His research interests include language testing, individual differences, and quantitative research methods. chris.pirotto@fukui-ut.ac.jp*

Robert Dykes *has been living and teaching English in Japan for over a decade. He holds a master's degree in Applied Linguistics and TESOL. His main research focus is language learner motivation and anxiety. robertd@jindai.ac.jp*

Received: November 4, 2018
Accepted: April 23, 2019

Appendix

Glossary of Terms (adapted with permission from Larson-Hall, 2016):

ANOVA: A model that includes at least one categorical independent variable. We are basically interested in seeing whether groups defined by the independent variable or variables performed differently on the dependent measure.

Assumption: When you use a statistical test there are certain assumptions that are made about your data...If your data do not meet the assumptions of the test, then you will have less power to find the true results.

Cohen's *d* (d): An effect size that measures the difference between two independent sample means. This is a group difference index of effect size. Cohen's *d* can start from zero and range as high as it needs to, although a d =1, meaning the differences between groups are as large as one standard deviation, would generally be considered a large effect size.

Confidence Interval (CI): The range of values around a statistic such as the mean that defines the range where the true population value of the statistic will be found on repeated testing of the research question.

Correlation: A statistical test that measures the strength of a relationship between two variables...The higher the correlation positively or negatively, the stronger the relationship.

Effect Size: An effect size measures how much effect can be attributed to the influence of an independent variable on a dependent variable, or to the relationship between variables. Effect sizes do not depend on sample size. Basically, effect sizes tell you how important your statistical results is.

Mean (X or M): The average of a group of numbers.

Null Hypothesis Significance Testing: An approach to testing statistics that features a null hypothesis and estimation of the conditional probability of the data with statistical tests. It focuses on p-values as the important criterion to use to determine whether the null hypothesis can be rejected or not.

Pearson's Correlation (r): The technical name for a correlation test that is parametric and inferential; correlation formally tests for a covariance in scores between two interval-level variables. The test asks whether there is a relationship between two variables.

Standard Deviation (usually abbreviated as s, s.d. or SD): A measure of how tightly or how loosely data are clustered around the mean.

***t*-test:** A parametric test that is used when you have one independent variable with only two levels and one dependent variable. You want to know if the two groups are different from each other.