Research Forum

Significance Testing, Research Quality, and Second Language Research: A Reflection and Review

Imogen Custance
Osaka Jogakuin College and University

In any article reporting on quantitative research, you are likely to find the letter *p*. This letter, or rather what follows it, can draw the eye as a busy researcher seeks to decide whether the results presented are of use. Yet the desire to use this shortcut belies a history of problems. Though the field of second language research has made progress in moving away from this all-or-nothing, significant-or-not fixation, improving awareness of issues with statistical techniques is necessary. This article reviews some issues with significance testing to raise or reignite awareness in this commonly used statistic.

定量調査を報告する論文では、「p」という文字を頻繁に目にする。この文字の後に続く数字は、多忙な研究者が提示された研究結果が有用かどうかを判断しようとする際に、特に注目される。しかし、このショートカットを使いたいという願望には、問題の歴史が潜んでいる。第二言語研究の分野では、このようなオール・オア・ナッシング、有意か無か、といった2元的な固定観念から脱却しつつあるが、統計的手法に関する問題意識を向上させることは必要である。本稿では、有意性検定に関するいくつかの問題を検討し、この広く使用されている統計量に対する認識を高める、あるいは喚起させることを目的とする。

Keywords: methodology; research methods; research quality; state of the field

https://doi.org/10.37546/JALTJJ47.1-4

In 2016, the American Statistical Association took the unprecedented step of publishing a statement on the use of *p* values in research (Wasserstein & Lazar, 2016). This step was taken in response to "highly visible discussions" (p. 129) regarding the use of null hypothesis significance testing (NHST) in a wide range of fields and to "draw renewed and vigorous attention to changing the practice of science with regards to the use of statistical inference" (p. 130). Yet as part of the statement they also emphasised that nothing mentioned was new information and that "statisticians and others have been sounding the alarm" about issues related to significance testing "for decades, to little avail" (p. 130). Indeed, over fifty years ago Bakan (1966) was already stating that the arguments in his paper were "hardly original" (p. 423) and in the following decade Carver suggested "educational research would be better off if it stopped testing its results for statistical significance" (1978, p. 378).

The aim of this article is not to call for an end to the use of statistical significance testing within second language research. Rather, it is to critically evaluate its use within the field and demonstrate the positive impact that greater consideration of issues associated with significance testing has had in what is hoped to be an accessible way. To this end, background information about the nature of statistical significance testing and some of the issues associated with it are introduced first. This is followed by a more detailed examination of its use in second language research and the impact a reliance on statistical significance testing as the main method for conducting quantitative analysis has had on the field. Next is an acknowledgement of the efforts being made to address these issues and the resulting changes in research quality, but also a recognition that increasing use of more advanced statistical techniques necessitates careful consideration of how research is used and by whom. I conclude that a shift in mindset regarding statistical significance testing is more appropriate than calling for the cessation of its use.

Background

Statistics has an influence on almost every aspect of modern life (Hand, 2008). It is used to make decisions and predictions about the future; to try and elucidate the relationships that underlie our reality. At its core, statistics is about modelling the world around us in such a way that we can understand it better. A model can never, however, be a perfect representation of reality. As such, there must be an inherent acceptance of uncertainty in any statistical model and in the results of a statistical test. Understanding the impact that this uncertainty has on the interpretation of results is necessary for those who wish to use statistics appropriately.

Null hypothesis significance testing (NHST) is one way in which uncertainty is acknowledged. In this type of testing, an assumption is made and termed the *null hypothesis*, H_0. This assumption is usually taken to be that there is no difference between two groups, with the actual idea of interest being that something has, in fact, caused there to be a difference between them. NHST is then used to check if the data that have been observed would be consistent if the null hypothesis were true. If the probability of the data occurring given the null hypothesis is true is below a predetermined alpha level α, usually α =.05 or α = .01, the null hypothesis is rejected. The probability is termed the *p* value, and "describes the probability that we would observe the value of the test statistic as extreme or more extreme than that actually observed, if the null hypothesis were true" (Hand, 2008, p. 89).

When calculating statistical significance using NHST, researchers must be careful to acknowledge the two possible mistakes that could arise when examining the *p* value and deciding whether to reject the null hypothesis. The first, a Type I error, is rejecting the null hypothesis when it is in fact true. The second, a Type II error, is failing to reject the null hypothesis when an alternate hypothesis is true. Hand (2008) explained these two types of error using the example of a court of law in which the null hypothesis is the assumption of innocence, with a Type I error the equivalent of an innocent person being found guilty, and a Type II error the equivalent of someone who is actually guilty being declared innocent. Norris (2015) highlighted that NHST, because only the null hypothesis is considered, will always help to avoid a Type I error—with a small enough alpha level, we can be relatively certain that the null hypothesis should be rejected. However, reducing the chance of a Type II error requires careful consideration of the statistical power necessary for a study. Field (2017) defines power as the "the ability of a test to find an effect" when there is an effect to be found and depends on the size of the effect, the sample size, and alpha-level or corrected alpha-level if multiple tests are conducted (p. 84).

While accepting that the *p* value can be useful because it helps researchers "be cautious in claiming that a difference or relationship they have observed in their sample data is actually rare…in comparison with the assumption that there is no such pattern" (Norris, 2015, p. 100), Norris is very critical of its use, in part because it is often misinterpreted as doing much more than this. Carver (1978) presented three "fantasies" (p. 383) about what information *p* values provide. These are misinterpreting the *p* value as the probability that the results were due to chance, the probability that a replication of the study would achieve the same result, and the probability that the research hypothesis is true. Carver, along with others (e.g., Field, 2017; MacInnes, 2022), also emphasised the importance of understanding that the probability of the null hypothesis being true given the data, $p(H_0|D)$, cannot be inferred from the probability of the result of the NHST, which is the probability of the data given the null hypothesis is true, $p(D|H_0)$. Other misinterpretations about *p* values raised by Greenland et al. (2016) include their ability to demonstrate whether a particular hypothesis is true or false, that the size of the *p* value itself is indicative of how strong the evidence for/against the null hypothesis is, and that finding statistical significance indicates an important discovery or observation has been made.

The reporting of effect sizes, which are standardized measures of how large or seemingly important a difference that has been identified is, has been advocated as a way to move away from a focus on NHST. Effect sizes not only give a measure of the importance of a discovery, but are also unaffected by sample size (Field, 2017). This is important because *p* values depend on the size of the sample (Field, 2017; Norris, 2015; Plonsky, 2015) and given a large enough sample size, it will always be possible to find a statistically significant difference between populations (Bakan, 1966). This means that whether a result is considered significant or not can be a function of the sample size, rather than the existence of an actual, meaningful difference or effect between groups. An effect size reported with a confidence interval provides more meaningful information about a result than a *p* value (Field, 2017; Norris & Ortega, 2000; Plonsky, 2015).

Misinterpretation and/or a lack of understanding of *p* values is an issue, but one that could potentially be solved through increased education. Indeed, the purpose of Greenland et al.'s (2016) paper was to provide a resource to help researchers "avoid and spot misinterpretations" (p. 337). However, researchers in a range of fields have gone beyond calling for increased understanding of *p* values and instead suggested that the use of NHST should be actively discouraged or stopped, with some journals banning its use (Trafimow & Marks, 2015). In the following section, I present some of the reasons given for such proposals, with reference to the use of NHST in second language research.

Issues With NHST in Second Language Research

It has been argued (e.g., Labaree, 2011; Porter, 1996) that professions with more applied and practical purposes "find themselves subject to the greatest external pressures and the strongest need to demonstrate the credibility of their claims through quantitative means" (Labaree, 2010, p. 624). The use of experimental design and quantitative analysis of the results is strongly associated with the scientific method, which in turn carries positive connotations of objectivity and trustworthiness. This might in part explain the prevalence of NHST in a range of fields, including second language research, despite the limited nature of the information provided by such tests. The use of quantitative research methods and an apparently objective method of determining the significance of claims can be seen to help legitimize a field which can in turn promote investment and development. Norris (2015) suggested that "The simplicity and apparent certainty of significance testing is alluring" (p. 101), which has made it a popular way to obtain and interpret quantitative research results. Field (2017) similarly suggested that "NHST seems to provide an easy way to disentangle the 'correct' conclusion from the 'incorrect' one" and that it is "appealing to teach" because students "can follow the rule that a $p < 0.05$ is 'significant' and a $p > 0.05$ is not" even if they do not understand the underlying logic of the test (p. 97).

Yet it has been suggested that the emphasis placed on achieving statistical significance in research results has in fact moved fields such as educational research (Carver, 1978), psychology (Chambers, 2017), and second language research (Plonsky, 2015) away from the ideals of the scientific method and scientific rigor. One of Carver's (1978) central arguments for stopping testing results for statistical significance is that it has led to an overemphasis on the finding of statistical, as opposed to scientific, significance in results. This is an argument echoed by Plonsky (2013, 2015), who further suggested that the tendency to not report non-significant results belies an underlying fixation on achieving statistical significance rather than examining what can be understood from the data collected.

A fixation on achieving statistical significance is problematic for a variety of reasons. First, the commonly used cut-off for having obtained a "significant" result, $p < 0.05$, is arbitrary (Plonsky, 2015) and "encourages all-or-nothing thinking" (Field, 2017, p. 99). Rather than considering the actual size of an effect, NHST encourages a knee-jerk reaction as to whether the data should be examined in more detail or not, or even reported at all. Second, it is likely that there is a bias towards publishing results that are found to be statistically significant (Marsden et al., 2018). According to Norris (2015), there "seems to be an artificial imbalance" (p. 104) between the number of studies finding statistical significance and those failing to within L2 research. Whether this is due to researchers deciding not to submit a study where no statistical significance was found, or journal editors rejecting it for perceived lack of importance, these results are "put away" leading to a skewed understanding of the research domain (Norris & Ortega, 2006, p. 21). Finally, the apparent preference for publishing positive, exciting results might lead researchers to engage in questionable research practices in order to have their work published (Chambers, 2017; Field, 2017; Marsden et al., 2018). This includes practices such as *p*-hacking, where researchers make decisions regarding what data or analyses to use, and which results to report, based on whether they yield statistically significant results, and HARKing (hypothesizing after the results are known), where a hypothesis is presented as having been made before data was collected or analysed when this was not the case (Field, 2017).

The aim of research is to understand and develop theories that explain the observations that are made. Research results can provide evidence for new theories; they can support or contradict previous findings. However, the use of NHST as a gatekeeper to publication is "a corrupt form of the scientific method" (Carver, 1978, p. 378). When publication is determined, or thought to be determined, by whether or not statistical significance is found, the process of theory falsification that is central to the scientific process falls apart. A negative result is considered "trash" (Plonsky, 2015, p. 24), is less likely to be published, and will not have an impact on the development of theory or future research. This can lead to the existence of *undead theories* (Ferguson & Heene, 2012), theories that continue to be used because negative evidence against them remains unpublished, or that otherwise resist attempts at falsification. An unexpected result of no statistical significance does not mean a study is necessarily uninformative or unimportant, or that it should not be considered for publication.

The presence of publication bias and an apparent disregard for results of non-significance started to receive more attention as L2 researchers began to look at conducting meta-analyses of primary research (e.g., Norris & Ortega, 2000). Meta-analyses are a powerful means of understanding the findings in a field of research. Whereas the findings of individual studies might be "attributable to chance variability as well as idiosyncrasies in design, analysis, sampling error [and] research setting" (Norris & Ortega, 2000, p. 423), a secondary analysis combining the results of several primary studies can help bring to light overall patterns that are applicable beyond the setting of any individual study. Publication bias is problematic for meta-analyses for the same reasons it is problematic to the field as a whole—if non-significant results are not published, they cannot be included in meta-analyses, and any patterns discerned in the meta-analyses will not reflect the reality that some studies have not found significance. There is also an issue when statistics, especially standard deviations or effect sizes, are not reported for non-significant results, a tendency identified by Plonsky (2013). The way in which meta-analyses are generally conducted is by finding primary data related to a treatment or condition and estimating the overall magnitude of any observed relationship or effect across the different studies. This involves an examination of the effect sizes in the primary studies. However, if the effect sizes or data required for calculating effect sizes, i.e., standard deviations, are not reported, it is not possible to include a study in a meta-analysis. Thus, a failure to report non-significant results or report all relevant statistics creates a significant barrier to the undertaking of meta-analytic research and the furtherance of the field through secondary research. Many of the issues with NHST mentioned so far are also underscored by Norris and Ortega (2000).

Reporting of results and questionable research practices are not the only issues associated with NHST in the field. Plonsky (2013, 2014) highlighted issues with the appropriacy of some of the statistical tests used in L2 research. He reported that the most commonly used inferential statistical test within the field from 1990–2010 was ANOVA, with 56% of studies using this type of analysis. Plonsky (2013, 2014) suggested that using a means-based test like ANOVA is problematic because they can obscure some of the information that is of interest to the field. When conducting an ANOVA, the means for the different groups in the analysis are taken and compared. By taking the mean, information relating to variance between those within the group is necessarily lost. Plonsky argued that given the complex nature of the constructs involved in L2 research, failing to preserve this variance reduces the conceptual validity of results obtained with these types of tests—comparing means is a practice that “sacrifices variance, informational richness, and statistical power for an analytic model that appears more straightforward” (Plonsky, 2014, p. 453).

A further issue with means-based testing is that researchers often conduct multiple tests which has a “debilitating effect on statistical power” (Plonsky, 2014, p. 453) as the alpha value must be adjusted for each additional test conducted. Indeed, low statistical power is thought to be a major concern for the field for a number of reasons (Lindstromberg, 2023; Plonsky, 2013). Statistical power is affected by both sample size and the expected effect size. As such, a study is likely to be underpowered if the sample sizes and/or the effect sizes is small. This is an issue with second language research because sample sizes are typically small (Lindstromberg, 2016; Plonsky, 2013). When a study is underpowered, the risk of a Type II error, failing to find a significant effect when there is one, is increased. Plonsky (2015) suggested that a power level of 0.8, or an 80% chance of detecting a real effect, is appropriate for social science research. However, in his study of research quality, Plonsky (2013) estimated that statistical power in the field, based on the median sample and effect sizes of the publications examined, was just 57% on average. In addition, research in the related fields of psychology and education have suggested that the effect sizes found in second language research are likely to be overestimated as a result of publication bias (Lindstromberg, 2023). As such, researchers might require even larger sample sizes to achieve sufficient statistical power when using NHST.

In sum, the number of issues related to statistical significance testing in second language research is indicative that it “is probably not well-conceived or accurately interpreted” (Norris, 2015, p. 106) and has resulted in a wide range of problems for the field.

The Impact of Publication Requirements

Despite the issues mentioned in the previous section, NHST continues to be used. However, changes in editorial policies and publication guidelines from top journals have started to deemphasise NHST in favour of procedures that highlight the scientific significance of results and academic rigour. The most recent American Psychological Association (APA) guidelines (American Psychological Association, 2019) contain a chapter focused on journal article reporting standards that details what information should be included for different types of research. These standards, initially developed in 2008 for quantitative research (APA Publications and Communications Board Working Group on Journal Article Reporting Standards, 2008) and revised in 2018 (Appelbaum et al., 2018), include the reporting of intended sample size and the statistical power analysis used to determine it, descriptive statistics, effect sizes, and exact *p*-values for all statistical tests whether a significant effect has been found or not. When establishing the initial standards in 2008, the working group aimed to create guidelines that would "[promote] sufficient and transparent descriptions of how a study was conducted and what researcher(s) found […to permit] the users of the evidence to judge more accurately the appropriate inferences and applications derivable from research findings" (APA Publications and Communications Board Working Group on Journal Article Reporting Standards, 2008, p. 847). They also highlighted that the suggested standards could encourage researchers to consider study plans more carefully, facilitate replication studies, and increase the number of studies that can be included in meta-analyses. The development of explicit reporting standards based on these ideas, in addition to calls for increased use of open science practices (e.g., Al-Hoorie et al., 2024; Liu et al., 2023; Marsden & Morgan-Short, 2023), demonstrate a desire to acknowledge and overcome some of the issues listed in the previous section.

One way to see the extent to which changes to publishing standards have impacted the field is to revisit articles published before they were introduced and consider how the research might be done differently if conducted now. A journal in second language research at the forefront of some guideline changes is *Language Learning,* which has required the reporting of effect sizes since 2000 (Ellis, 2000), specified a range of guidelines for reporting quantitative research in 2015 (Norris et al., 2015), and introduced registered reports, whereby a study can essentially be approved for publication before results are known in 2018 (Marsden et al., 2018). In 1999, the editor's statement highlighted that one of the journal's strengths was the "high quality of its empirical research" and "focus on the systematic collection, analysis and evaluation of data" (Ellis, 1999, p. vi). In the same volume, an article, Skehan and Foster (1999), reporting the results of an experiment examining the effect of task structure and processing load, was published.

In their study, Skehan and Foster used a 2 x 4 between-subjects design to investigate two tasks and four performance conditions. The nature of the tasks and measures used, including specific reasons for why certain measures were not used are explained, and descriptive statistics were reported along with group sample sizes. Participants were randomly assigned to task and condition, and measures were checked for collinearity. There is probably sufficient detail included within the article to make replication of the study possible, as required by the editor (Ellis, 1999).

However, only 47 participants were involved in the study and whether the study had enough statistical power for the two-way ANOVAs conducted should be considered. Using the Plonsky and Oswald (2014) suggested benchmarks for interpreting effect sizes (Cohen's *d*) for L2 research of small ($d = 0.4$), medium ($d = 0.7$), and large ($d = 1.0$) it is possible to calculate the sample size necessary to achieve effect sizes of these levels. I used the computer program G*Power (Faul et al., 2009). The alpha level was set to .05, and power to .8, the level suggested for social science research (Plonsky, 2015). G*Power calculates the *N* size based on a Cohen's *f*, so I converted the *d* values suggested by Plonsky and Oswald (2014) using the formula

$$f^2 = \frac{d^2}{2k}$$

(Statistics How To, n.d.) where *k* is the number of groups ($k = 8$ for Skehan and Foster's study). Table 1 shows the effect sizes (*f* and *d* values), and the *N* sizes suggested to be necessary if these effect sizes were expected. They indicate that even if a large effect size was expected, the study would have been underpowered given the actual sample size ($N = 47$).

Table 1

Estimation of Necessary N-Sizes for Skehan and Foster's (1999) Study

Effect Size	Cohen's *d*	Cohen's *f*	*N* size
Small	0.4	0.100	1095
Medium	0.7	0.175	360
Large	1.0	0.250	179

An early consideration of statistical power within Skehan and Foster's study design process might have helped the researchers to adjust their research design. The researchers could have examined fewer performance conditions, a strategy recommended by Norris and Ortega (2000), or chosen to analyze the data differently. Alternatively, they might have decided to conduct additional data collection to achieve a more appropriate *N* size. In either case, the study would likely have been improved, highlighting why addressing intended sample size is beneficial to research.

There are also differences in what and how the statistical results might be reported today. First, no information was given about the results of assumption checks on the data. Though the researchers might have conducted these, by not reporting the results, readers cannot judge whether the *p*-values presented are accurate. At the time the article was published, ANOVA was considered a robust test, so issues with assumption violations might not have been considered, but this would not be the case now. Second, even if the data met the assumptions required to produce accurate *p*-values, no effect sizes were reported, nor any focus given to how meaningful the statistically significant differences between tasks and performance conditions were. The arguments made in the discussion would have been strengthened were they supported by effect size information. Finally, visualizations of the data, including effect sizes and confidence intervals, would potentially have made interpretation of the results simpler.

Where Do We Stand?

To an extent, by not stopping testing results for statistical significance when the issues with NHST were first raised, "a great deal of mischief" (Bakan, 1966, p. 423) and damage has probably been wrought in various fields, including that of second language research. NHST has become the "go-to analytic approach…for making sense of numerical data" (Norris, 2015, p. 97) as a result of a self-fulfilling cycle whereby it is "[taught] because it's what we do; …[done] because it's what we teach" (Wasserstein & Lazar, 2016, p. 129). The misuse and misinterpretation of *p* values over the years has likely resulted in a range of somewhat erroneous theories gaining traction while the failure to report non-significance or details related to non-significant results has harmed the field's ability to conduct meta-analyses. However, there are clear indications that the issues raised all those years ago are now being addressed much more proactively, as exemplified by the various changes to editorial policies within second language research journals, the publication of books focused on the use of statistics within L2 research, the increase in research that has been conducted into study quality, and the launch of a journal, *Research Methods in Applied Linguistics*, that is "devoted exclusively to the study and advancement of methods and approaches in language-related research" (Li & Prior, 2022, p. 1). The shift from a focus on a significant result to ensuring more transparent reporting practices and how a finding might impact our understanding of theory is positive.

However, the full impact of these calls for change must not be underestimated. The field of second language research is maturing, as partly demonstrated by the increased use of more advanced statistical techniques. Khany and Tazik (2019) found that the number of research articles requiring the knowledge of intermediate or advanced techniques increased from 20.61% between 1986–1995 to 39.08% between 2006–2015. While this is far from a negative development, it necessitates an examination of how this research is used and by whom. Loewen et al. (2020) found "mixed evidence of…researchers' ability to use and interpret" (p. 883) statistical information and "limited overlap" (p. 884) between self-perceived statistical knowledge and actual knowledge as determined by a test. They suggested this is likely to make calls for increased rigour and use of more advanced techniques difficult to achieve. If the use of more advanced techniques is being advocated, it is necessary to improve statistical literacy which is "critical for both producers and consumers of L2 research" (Gonulal et al., 2017, p. 4). While courses that cover statistical techniques can be effective in raising learners' confidence in their ability to interpret and use such techniques (Gonulal et al., 2017), whether institutions are able to offer such courses, or professionals take them, will impact how well the field can adapt to the presence of increasingly complex statistics.

Most of the arguments from second language researchers presented in this paper are from those aligned with the field of second language acquisition. Gass et al. (2021) argued that the area of second language acquisition has "evolved into a unique area of inquiry seeking to understand how second languages are learned without an emphasis on how they are taught" (p. 247). However, the knowledge and theory produced in this field continues to have an impact on that of language teaching as "the two disciplines feed one another and are relevant to one another" (Gass et al., 2021, p. 247). Not considering how to maintain the accessibility of this research to members of the overarching field of second language research has the potential to cause a different type of damage to that resulting from misuse of *p* values. The use of *p* values can make the results of a study more accessible to other L2 researchers. Effectively, by including *p* values, the reporting of results remains in line with what most researchers in the field would expect and find easy to interpret. In this way, continued reporting of *p* values makes it possible for more people to learn something from a publication. While there might be some misinterpretation of what the *p* value means, by keeping this avenue of interpretation open, the field itself remains more open and inclusive.

Of course, there are other ways in which knowledge and expertise develop. Just as Isaac Newton saw further by standing on the shoulders of giants, so does anyone in a field learn from what has gone, or been published, before. This is not limited to the results of a study, but also the methods and analyses employed to answer the research questions posed. Thus, greater attention to, and reporting of, whether assumptions are met before conducting a statistical test highlights the importance of this issue to researchers. The shift in focus towards not just the results, but the appropriacy of methods used is already evident within the field. The introduction of registered reports (e.g., Marsden et al., 2018) is another step in the direction of emphasizing study quality and confidence in results over how novel the results observed are. Yet these facts do not preclude the use of statistical significance testing. Rather, they emphasize that it is necessary to consider whether a given test is appropriate for the data obtained or relevant for answering the research questions posed.

Conclusion

In conclusion, and as many have said before (e.g., Bakan, 1966; Norris, 2015; Wasserstein & Lazar, 2016), statistical significance testing itself is not at fault; the problem lies with how it has been used and overemphasised. With this in mind, it is necessary that journals and journal reviewers take special care to check that publications do not risk propagating false ideas regarding these tests, and that researchers and graduates are educated to ensure that they understand the limitations of statistical significance. Brown's (2016) *Statistics Corner* is a collection of articles published in the JALT Testing and Evaluation SIG's publication that is a useful resource in this regard. In addition, when planning a study, researchers should carefully consider the statistical power necessary for conducting specific tests, and report descriptive statistics including means and standard deviations, effect sizes, and exact *p*-values for all results, including those that are non-significant. The field and research techniques used within it will continue to evolve and time will tell if NHST "survives as a useful technique or is replaced" (Norris, 2015, p. 123). For now, it continues to play a role and perhaps needs to as a simple, if somewhat limited, part of the research landscape.

Imogen Custance is a lecturer at Osaka Jogakuin University and College. Her research interests include the development of speaking skills and communicative competence.

References

Al-Hoorie, A. H., Cinaglia, C., Hiver, P., Huensch, A., Isbell, D. R., Leung, C., & Sudina, E. (2024). Open science: Considerations and issues for TESOL research. *TESOL Quarterly*, *58*(1), 537–556. https://doi.org/10.1002/tesq.3304

American Psychological Association. (2019). *Publication manual of the American Psychological Association* (7th ed.).

APA Publications and Communications Board Working Group on Journal Article Reporting Standards. (2008). Reporting standards for research in psychology: Why do we need them? What might they be? *American Psychologist*, *63*(9), 839–851. https://doi.org/10.1037/0003-066X.63.9.839

Appelbaum, M., Cooper, H., Kline, R. B., Mayo-Wilson, E., Nezu, A. M., & Rao, S. M. (2018). Journal article reporting standards for quantitative research in psychology: The APA Publications and Communications Board task force report. *American Psychologist*, *73*(1), 3–25. https://doi.org/10.1037/amp0000191

Bakan, D. (1966). The test of significance in psychological research. *Psychological Bulletin*, *66*(6), 423–437. https://doi.org/10.1037/h0020412

Brown, J. D. (2016). *Statistics corner: Questions and answers about language testing statistics*. TEVAL SIG.

Carver, R. (1978). The case against statistical significance testing. *Harvard Educational Review*, *48*(3), 378–399. https://doi.org/10.17763/haer.48.3.t490261645281841

Chambers, C. (2017). *The seven deadly sins of psychology: A manifesto for reforming the culture of scientific practice.* Princeton University Press. https://doi.org/10.1515/9781400884940

Ellis, N. C. (1999). Editor's statement. *Language Learning*, *49*(1), v–vi. https://doi.org/10.1111/1467-9922.00068

Ellis, N. C. (2000). Editor's statement. *Language Learning*, *50*(3), xi–xiii. https://doi.org/10.1111/0023-8333.00135

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, *41*(4), 1149–1160. https://doi.org/10.3758/BRM.41.4.1149

Ferguson, C. J., & Heene, M. (2012). A vast graveyard of undead theories: Publication bias and psychological science's aversion to the null. *Perspectives on Psychological Science*, *7*(6), 555–561. https://doi.org/10.1177/1745691612459059

Field, A. (2017). *Discovering statistics using IBM SPSS statistics* (5th ed.). Sage.

Gass, S., Loewen, S., & Plonsky, L. (2021). Coming of age: The past, present, and future of quantitative SLA research. *Language Teaching*, *54*(2), 245–258. https://doi.org/10.1017/S0261444819000430

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

Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. B., Poole, C., Goodman, S. N., & Altman, D. G. (2016). Statistical tests, *p*-values, confidence intervals, and power: A guide to misinterpretations. *European Journal of Epidemiology*, *31*(4), 337–350. https://doi.org/10.1007/s10654-016-0149-3

Hand, D. J. (2008). *Statistics: A very short introduction*. OUP Oxford. https://doi.org/10.1093/actrade/9780199233564.001.0001

Khany, R., & Tazik, K. (2019). Levels of statistical use in applied linguistics research articles: From 1986 to 2015. *Journal of Quantitative Linguistics*, *26*(1), 48–65. https://doi.org/10.1080/09296174.2017.1421498

Labaree, D. (2011). The lure of statistics for educational researchers. *Educational Theory*, *61*(6), 621–632. https://doi.org/10.1007/978-90-481-9873-3_2

Li, S., & Prior, M. T. (2022). Research methods in applied linguistics: A methodological imperative. *Research Methods in Applied Linguistics*, *1*(1), 100008. https://doi.org/10.1016/j.rmal.2022.100008

Lindstromberg, S. (2016). Inferential statistics in Language Teaching Research: A review and ways forward. *Language Teaching Research*, *20*(6), 741–768. https://doi.org/10.1177/1362168816649979

Lindstromberg, S. (2023). The winner's curse and related perils of low statistical power − spelled out and illustrated. *Research Methods in Applied Linguistics*, *2*(3), 100059. https://doi.org/10.1016/j.rmal.2023.100059

Liu, M., Chong, S. W., Marsden, E., McManus, K., Morgan-Short, K., Al-Hoorie, A. H., Plonsky, L., Bolibaugh, C., Hiver, P., Winke, P., Huensch, A., & Hui, B. (2023). Open scholarship in applied linguistics: What, why, and how. *Language Teaching*, *56*(3), 432–437. https://doi.org/10.1017/S0261444822000349

Loewen, S., Gönülal, T., Isbell, D. R., Ballard, L., Crowther, D., Lim, J., Maloney, J., & Tigchelaar, M. (2020). How knowledgeable are applied linguistics and SLA researches about basic statistics?: Data from North American and Europe. *Studies in Second Language Acquisition*, *42*(4), 871–890. https://doi.org/10.1017/S0272263119000548

MacInnes, J. (2022). *Statistical inference and probability*. Sage. https://doi.org/10.4135/9781529682748

Marsden, E., & Morgan-Short, K. (2023). (Why) Are open research practices the future for the study of language learning? *Language Learning, 73*(2), 344–387. https://doi.org/10.1111/lang.12568

Marsden, E., Morgan-Short, K., Trofimovich, P., & Ellis, N. C. (2018). Introducing registered reports at Language Learning: Promoting transparency, replication, and a synthetic ethic in the language sciences. *Language Learning*, *68*(2), 309–320. https://doi.org/10.1111/lang.12284

Norris, J. M. (2015). Statistical significance testing in second language research: Basic problems and suggestions for reform. *Language Learning*, *65*(S1), 97–126. https://doi.org/10.1111/lang.12114

Norris, J. M., & Ortega, L. (2000). Effectiveness of L2 instruction: A research synthesis and quantitative meta-analysis. *Language Learning*, *50*(3), 417–528. https://doi.org/10.1111/0023-8333.00136

Norris, J. M., & Ortega, L. (2006). The value and practice of research synthesis for language learning and teaching. In J. M. Norris & L. Ortega (Eds.), *Synthesizing research on language learning and teaching* (pp. 1–50). John Benjamins. https://doi.org/10.1075/lllt.13.04nor

Norris, J. M., Plonsky, L., Ross, S. J., & 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. *The Modern Language Journal*, *98*(1), 450–470. https://doi.org/10.1111/j.1540-4781.2014.12058.x

Plonsky, L. (2015). Statistical power, *p* values, descriptive statistics, and effect sizes. In L. Plonsky (Ed.), *Advancing quantitative methods in second language research* (1st ed., pp. 23–45). Routledge. https://doi.org/10.4324/9781315870908-3

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

Porter, T. M. (1996). *Trust in numbers: The pursuit of objectivity in science and public life*. Princeton University Press. https://doi.org/10.1515/9780691210544

Skehan, P., & Foster, P. (1999). The influence of task structure and processing conditions on narrative retellings. *Language Learning*, *49*(1), 93–120. https://doi.org/10.1111/1467-9922.00071

Statistics How To. (n.d.). *Cohen's f statistic: Definition, formulas*. Retrieved 10 August 2024, from https://www.statisticshowto.com/cohens-f-statistic-definition-formulas/

Trafimow, D., & Marks, M. (2015). Editorial. *Basic and Applied Social Psychology*, *37*(1), 1–2. https://doi.org/10.1080/01973533.2015.1012991

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA statement on *p*-values: Context, process, and purpose. *The American Statistician*, *70*(2), 129–133. https://doi.org/10.1080/00031305.2016.1154108