Advancing quantitative research methods

Shawn Loewen and Aline Godfroid

Quantitative data analysis has a strong tradition in applied linguistics research (e.g., Loewen & Gass, 2009; Norris, Ross, & Schoonen, 2015). Even though the amount of qualitative research has been increasing recently, the majority of published applied linguistics studies still employ quantitative analysis (Khany & Tazik, 2018). Quantitative methods give applied linguistics researchers the tools to describe their data and examine whether trends generalize beyond the language learners, contexts, and tasks included in a particular study. As such, quantitative methods are a cornerstone of many types of applied linguistics research. Indeed, some researchers are adopting even more sophisticated quantitative analysis techniques in their research.

Recently, there has been an increased level of scrutiny of the quantitative research methods and reporting practices employed by applied linguists. Byrnes (2013) went so far as to call it a “methodological turn” (p. 825). Evidence for this focus comes from a variety of sources, including the increasing number of discipline-specific books pertaining to quantitative analysis (e.g., Gries, 2013; Larson-Hall, 2016; Loewen & Plonsky, 2016; Lowie & Seton, 2013; Plonsky, 2015), special journal issues (Norris et al., 2015 in Language Learning; Li, Shintani, & Ellis, 2015 in Applied Linguistics), and individual journal articles (e.g., Al-Hoorie & Vitta, 2018; Godfroid & Spino, 2015; Khany & Tazik, 2018; Lindstromberg, 2016b; Marsden, Morgan-Short, Thompson, & Abugaber, 2018; Mizumoto & Plonsky, 2016; Plonsky, 2013, 2014).

This chapter will explore the current discussion around the need for improvement in the use of quantitative research methods by applied linguists. The chapter will focus on (a) current practices, (b) null hypothesis significance testing, (c) the open science movement, (d) recommendations for continued advancement of quantitative research methods, and (e) statistical literacy and training in applied linguistics programs.

Critique of current practices

Concern with quantitative research methods in applied linguistics and its related fields is not a new phenomenon. For example, Hatch and Farhady (1982) published a research manual specifically for applied linguists. A decade later, a revised version of the book appeared (Hatch & Lazaraton, 1991). However, for numerous years, these were the only discipline-specific books
available for quantitative research. Additionally, there were a few journal articles that dealt specifically with statistical analyses. Specifically, J. D. Brown published several articles that provided information about statistical analysis. One brief piece (Brown, 1990) advised researchers against the use of multiple t-tests in their analyses, while two articles explored “statistics as a foreign language” (Brown, 1991, 1992).

However, it was probably the publication of Norris and Ortega’s (2000) meta-analysis of the effectiveness of L2 instruction that was a catalyst for the current, intense interest in quantitative methods. In their meta-analysis, Norris and Ortega examined effect sizes from 49 unique studies; however, they had to exclude 42% of potential studies due to a lack of full reporting of the statistics necessary for meta-analysis. In addition to presenting the results of the meta-analysis regarding the effectiveness of L2 instruction, Norris and Ortega provided a relatively lengthy critique of the ways in which quantitative analyses were conducted and reported in the studies they investigated. They identified issues such as a failure to report basic descriptive and inferential statistics and a reliance on null hypothesis significance testing rather than effect sizes. Consequently, Norris and Ortega called for improvement in the way that researchers conducted and reported quantitative studies.

As the research synthesis trend grew, authors of new meta-analyses found that they had to exclude an unfortunately large number of studies due to insufficient reporting practices (e.g., Plonsky, 2011; Qureshi, 2016; Russell & Spada, 2006; Yan, Maeda, Lv, & Ginther, 2016; Ziegler, 2016). Thus, an important result of these studies was the call for better reporting practices, including reporting all means and standard deviations as well as effect sizes and p values regardless of which side of the alpha level they were on.

Additional exhortations for better reporting practices came from journal editors who more clearly stipulated the information that needed to be included in quantitative submissions to their journals. For example, Ellis (2000) made it mandatory to report effect sizes for Language Learning. Currently, many of the top applied linguistics journals have specific information about conducting quantitative research and reporting descriptive and inferential statistics (e.g., Lindstromberg, 2016a for Language Teaching Research; Malboob et al., 2016 for TESOL Quarterly; Norris, Plonsky, Ross, & Schoonen, 2015 for Language Learning).

**Alternatives and supplements to Null Hypothesis Significance Testing**

Related to the lack of effect size reporting is the criticism in applied linguistics quantitative research of null hypothesis significance testing (NHST) (Godfroid & Spino, 2015; Norris, 2015). In fact, this aversion to NHST mirrors similar concerns in the larger sciences as well. A primary criticism is that finding out whether or not a specific difference between two samples is ‘significant’ or not is not very helpful. Because NHST technically allows only a dichotomous answer – significant or not significant – it does not provide any information about the strength of an effect. For example, a t-test returns a p value that is either over or under the predetermined alpha level (typically .05). Technically any value below the alpha level is considered ‘significant’ regardless of whether it is .04 or .0001. Similarly, values over .05, be they .06 or .60 are reported as non-significant, though “surely, God loves the .06 nearly as much as the .05” (Rosnow & Rosenthal, 1989, p. 1277). But seeing the world dichotomistically – with statistical tests producing only significant or non-significant results – offers a false sense of security. In reality, p values perform a “dance of the p values” (Cumming, 2008): they vary widely from study to study, so that “NHST gives only poor information about the likely result of a replication” (Cumming, 2014, p. 13).
The shortcomings of NHST is where the emphasis on effect sizes and confidence intervals came in. Effect sizes can provide the magnitude of an effect, with larger effect sizes showing a stronger effect. Furthermore, confidence intervals (CIs) represent the precision of an estimate (e.g., an estimated effect size), with smaller CIs indicating more precision. Effect sizes and CIs invite researchers to think about the relative strength of the evidence provided by their data. Thus, a question that can be answered by NHST such as, does attention help language learning?, results in either a yes or no answer. However, a question such as to what extent does attention help language learning? can be answered with more precision with effect sizes and CIs. In fact, Plonsky and Oswald (2014) have come up with discipline-specific guidelines for interpreting effect sizes. What counts as a small, medium, or large effect in applied linguistics is substantially larger (in effect size units) than what counts as a similar effect in neighboring fields. While this finding in itself is informative, the authors also acknowledged that the larger effect sizes may reflect deeper issues in applied linguistics, such as the influence of small sample sizes and publication bias (also see Lindstromberg & Eyckmans, 2017).

Together with meta-analysis, effect sizes and CIs play a key role in the new statistics advocated by some quantitative researchers in psychology and applied linguistics (Cumming, 2012, 2014; Cumming & Calin-Jageman, 2017; Larson-Hall, 2016). Cumming, in particular, has advocated for abandoning NHST and p values in favor of magnitude estimation (i.e., effect sizes). Although applied linguists nowadays more commonly opt for a hybrid approach, and report p values along with effect sizes, dropping p values altogether would be a step forward in reducing publication bias caused by journals publishing only studies with statistically significant results.

Finally, for some questions, NHST may not be suitable at all, because traditional statistical tests are geared towards finding evidence of differences between treatments or groups (the alternative hypothesis), whereas sometimes it is the lack of a difference (the null hypothesis) that is of interest. When the goal is to show that two treatments or groups do not differ, equivalence tests (Godfroid & Spino, 2015) or Bayesian hypothesis testing (Dienes, 2014; Mackey & Ross, 2015; Norouzian, de Miranda, & Plonsky, 2019; Ross & Mackey, 2015) may be more appropriate. These approaches will be discussed in more detail in the section on new and sophisticated analyses.

The open science movement

In the spirit of enhanced study quality, researchers have called for the increased replication of applied linguistics studies (e.g., Marsden, this volume; Marsden, Morgan-Short, Thompson et al., 2018; Polio & Gass, 1997; Porte, 2012). A replication is a repetition, exact or approximate, of a previous study designed to test the reliability and generalizability of the study findings (Porte, 2012). Because quantitative research aims at generalization, it is important that findings be replicated, but unfortunately, researchers do not always conduct replication studies, and when they do, a replication of results does not always occur. For instance, an attempt to replicate 100 psychology papers resulted in about a 39% success rate (Open Science Framework, 2015). This result strengthened concerns about a “replication crisis” (Pashler & Wagenmakers, 2012) because it questioned the empirical basis on which theories in psychology are built. In second language acquisition (SLA), Marsden, Morgan-Short, Thompson et al. (2018) estimated only one in 400 studies self-identified as a replication. The authors did not evaluate the reproducibility of research findings in SLA because there were too many changes between the initial and replicated studies; however, they argued instead for more and better quality replications in SLA. Both multisite, multiauthor collaborations and small research teams can contribute to this goal, as demonstrated by Morgan-Short et al.’s (2018) and Lindstromberg and Eyckmans’ (2017) recent replication research.
In recent years, it has become easier to replicate studies in applied linguistics because of the IRIS (www.iris-database.org) repository (Marsden, Mackey, & Plonsky, 2016), a free, online database of research materials, data sets, and analysis protocols. Applied linguists who have developed testing instruments, treatment materials, or other research materials for published studies are encouraged to upload these materials to the IRIS repository. Likewise, IRIS now also accepts data sets, analysis protocols, and registered reports (discussed below) in an effort to promote open science practices. To date, IRIS contains more than 3,600 materials and analysis protocols (Marsden, Morgan-Short, Thompson et al., 2018), all freely available for use in primary research or training in research methods classes.

Registered reports, or the preregistration of study protocols, are another initiative aimed at promoting open science. These reports are a new type of publication, created specifically to counteract publication bias (i.e., favoring of studies with statistically significant findings). To achieve this goal, the review of a registered report takes place in two stages. The first-stage review, which is arguably the most critical, consists of a full peer review of the study’s rationale, methods, materials, and analysis protocol before the data have been collected. The second-stage review, which is conducted after the results are known, focuses only on whether researchers adhered to the approved protocol. Articles are thus judged on their scientific merit and methodological rigor and not on whether the results are significant or not. This makes registered reports a powerful tool for counteracting publication bias, allowing researchers to “focus on reporting a confirmatory analysis . . . , without the need to hunt for positive and clean results” (Nosek & Lakens, 2014, p. 138).

At the time of writing, a total of 112 journals from multiple disciplines had adopted registered reports (Center for Open Science, 2018). In applied linguistics, Language Learning and Bilingualism: Language and Cognition recently introduced registered reports as a new article category (Marsden, Morgan-Short, Trofimovich, & Ellis, 2018). To incentivize the preregistration of research, authors are awarded a Preregistered badge (one of three possible open science badges) upon successful publication of a preregistered study (Trofimovich & Ellis, 2015). The open science movement in applied linguistics is growing, as multiple journals, such as Language Learning, Studies in Second Language Acquisition, and The Modern Language Journal, are now offering open science badges. These badges are displayed prominently on an article’s title page when authors make their materials and/or data publicly available, and/or preregister their studies. This reward system aims to promote a collaborative and transparent research culture, where researchers have easy access to each other’s materials and data and are encouraged to replicate and build on each other’s work to advance knowledge in the field.

New and more sophisticated analyses

Although it is unclear if statistical analyses in applied linguistics are becoming more advanced in general, it is true that more sophisticated quantitative analyses are being used by some applied linguistics researchers (see Pfenninger & Neuser, this volume). For a long time, t-tests, ANOVA, and correlation have been the workhorse of applied linguistics research (Baffoe-Djan & Smith, this volume; Khany & Tazik, 2018; Lindstromberg, 2016b; Plonsky, 2013; Plonsky & Oswald, 2014) and to a large extent this may still be true today. At the same time, t-tests, ANOVA, and correlation are simple cases of what is known as the ‘linear model’ (an equation that describes the outcome). A growing number of applied linguists, many of them psycholinguists and corpus linguists (Gries, 2015), are adopting more advanced expressions of the linear model, such as linear mixed-effects models. Although it is unclear whether linear mixed-effects models will replace ANOVA, as is happening in psychology (Matuschek, Kliegl, Vasishth, Baayen, & Bates,
Linear mixed effects models offer the same flexibility as multivariate regression (Jeon, 2015; Plonsky & Ghanbar, 2018), but with the possibility of filtering out additional sources of variance. Compared to mean-based analyses such as ANOVA and regression, mixed-effects models have the more ambitious goal of predicting each individual observation (Baayen & Milin, 2010). Each observation is thus treated as a unique data point, with no averaging by participants or items. This method allows the model to take full advantage of the information in the data. Researchers, on their end, need to ensure they have a sufficient number of participants and items per condition so effects stand out as “fixed effects” amidst the trial-to-trial variation captured by the random effects (Godfroid, 2020).

As the field progresses from univariate to multivariate analyses, more discipline-specific literature is being published. For example, Plonsky’s (2015) edited volume specifically addressed advanced quantitative analysis as a set of techniques that can better capture the complex, multivariate nature of L2 learning and use. The volume included chapters on techniques such as cluster analysis (Staples & Biber, 2015), exploratory factor analysis (Loewen & Gonulal, 2015), structural equation modeling (Schoonen, 2015), discriminant analysis (Norris, 2015), and Rasch analysis (Knoch & McNamara, 2015). In addition to this edited volume, journal articles have been published on analyses such as exploratory factor analysis (Plonsky & Gonulal, 2015) and structural equation modeling (Winke, 2014).

In some cases, however, researchers may find that what they need is not more statistical sophistication but a different way of doing statistics. Consider the following example: a researcher wants to know if L2 learners negotiate more or differently with their partner when doing a problem solving task in face-to-face or online settings (e.g., Rouhshad, Wigglesworth, & Storch, 2015). NHST dictates that the researcher will assume no difference between the two settings as the null hypothesis, regardless of what her true beliefs are. Now imagine that the difference between the two settings was $t(33) = 1.21, p = .23, d = .28$. Can we conclude from this that online interaction is as effective as face-to-face interaction? The answer is no, because we did not actually assert this hypothesis (Dienes, 2014; Godfroid & Spino, 2015; Norris, 2015).

To adduce positive evidence for the no-difference hypothesis, alternative statistical approaches such as equivalence tests or Bayesian analysis are needed. Although these approaches differ fundamentally in their details, both require researchers to engage with past research (e.g., on face-to-face and online interaction) to come up with a range of possible values for the new study. In the case of equivalence testing, the researcher’s task is to define a region within which the groups’ performance is deemed equivalent, for instance $\leq 10\%$ difference in negotiation behavior (Godfroid & Spino, 2015). Bayesian hypothesis testing, on the other hand, requires researchers to specify a theory of the alternative hypothesis (Dienes, 2014). What researchers will get in return is a set of statistical results that will allow them to assess the evidence for difference as well as equivalence between conditions. Specifically, when traditional tests return a non-significant $p$ value, as in the earlier example, the new analyses can determine whether the two conditions are truly equivalent (in terms of interaction patterns) or whether a larger sample size would be needed to uncover true effects in the data. In this way, equivalence tests and Bayesian hypothesis testing allow researchers to make the most out of their non-significant results (for an example, see Kim & Godfroid, 2019).

Lastly, the movement towards greater analytical sophistication and innovation should not lead us to forget the basics of good statistical analysis. Data visualization, descriptive statistics, and CIs are essential tools in any quantitative researcher’s tool kit, regardless of what additional tests they plan to do. Highlighting this point, Larson-Hall (2017) presented
a historical overview of graphing practices in three applied linguistics journals from their origin up to 2011 or 2012. To make optimal use of journal article space, graphs ought to be *data accountable* or *data rich*. Data-accountable or data-rich graphs present all or most of the original data points and also show the larger trends, so readers can see “both the forest and the trees” (Larson-Hall, 2017, p. 264). A nice example of a data accountable graph is a beeswarm plot—a boxplot overlaid with individual data points.

Using quality visuals aligns well with the open science value of research transparency. Because quality visuals let readers view the data and evaluate trends for themselves, readers can form their own opinion of the presented findings. Similarly, adding CIs to visual representations may guide readers away from NHST thinking, following the new statistics. Finally, in some contexts, limited access to participants may make it challenging to recruit larger participant samples. In such situations, a rich description based on trends seen in quality visuals will be researchers’ primary tool and should be preferred to running inferential statistics on an underpowered data set.

**Statistical training**

One question that has been raised is how applied linguistics researchers go about acquiring their statistical knowledge. Several studies (e.g., Lazaraton, Riggenbach, & Ediger, 1987, Loewen et al., 2014) have investigated both professors’ and graduate students’ perceptions of their statistical knowledge. In addition, these studies inquired about statistical training. Loewen et al. (2014) found that roughly 80% of doctoral students and professors had taken at least one statistics course, with the median being two courses. However, only 30% of professors and 13% of doctoral students felt that their statistical training had been adequate. In an effort to discover more about what constitutes statistical training in applied linguistics, Gonulal, Loewen, and Plonsky (2017) conducted a self-report study of graduate students’ perceptions of their statistical knowledge and self-efficacy at the beginning and end of semester-long quantitative research methods courses. The researchers sampled graduate students from four different applied linguistics programs in North America and found that students reported statistically significant increases in their knowledge of common inferential statistics.

In an additional investigation, Gonulal (2016) used a statistical literacy assessment to investigate the basic statistical knowledge of 125 graduate students in applied linguistics programs in North American universities. Rather than relying on participants’ self-reported perceptions of their statistical literacy, Gonulal asked participants to respond to a 28-item assessment that asked participants to interpret statistical information. Gonulal found that predictors of statistical literacy included taking more statistics courses, doing more self-training, and conducting more quantitative research.

Finally, it seems that the emphasis on improved reporting practices has had at least some impact. In their survey of statistical literacy, Lazaraton et al. (1987) did not even include *effect size* in their list of statistical terms on their survey. However, 27 years later, Loewen et al. (2014) found that *effect size* rated an average score of 4.17 (SD = .175) out of 6 points when respondents were asked about their ability to interpret effect sizes. The term also grouped with other common inferential statistics.

**Recommendations for the future**

It is no exaggeration to say that these are exciting times to be a quantitative applied linguist. With ongoing reform movements and intense innovation efforts quantitative researchers have
many things to read and learn about. We invite the applied linguistics research community
to join these various initiatives, which bring the promise of a science that is more accessible,
more transparent, and ultimately – we believe – better. Under the broad theme of “open sci-
ence”, researchers should be open about all stages of their work, from study design, to data
collection and analysis, to study reporting. At the stage of study design, researchers can pre-
register their study so that the research questions and methodology are reviewed and registered
before data collection. For experimental materials and instruments, sharing means caring. By
uploading their materials, instruments, and analysis protocols to online repositories such as the
IRIS database (www.iris-database.org), researchers help cement a collaborative research ethic,
whereby different research teams all work on the same, larger questions. When writing up an
article, researchers should report all their findings, even if they do not align with their hypoth-
eses or are statistically non-significant. Having a study preregistered may make it easier for
researchers to do so and will ultimately help counteract publication bias in applied linguistics.
Hence, the overall idea is to conduct research in a transparent manner, so others can retrace
previous steps and replicate the work. Additionally, with such procedures, a study’s results can
be included in future meta-analyses.

At a personal level, researchers should invest in their statistical literacy. It pays to learn the
basics and attempt to run the best possible analysis every single time. Inspecting descriptives,
checking assumptions, and visualizing the data all sound fairly simple, but these steps are not
always done well or, worse, are sometimes skipped in the rush to get a p value. Researchers
are also encouraged to move beyond the basics and acquire skills that will benefit their par-
ticular line of work. When adding new techniques to their statistical repertoire, researchers
will find they can read research publications with more confidence and autonomy. Even so,
complexity of analysis should not translate to complexity of reporting, but instead, clarity
and completeness should come first. As future generations prepare to enter the field, they will
benefit from seeing these best practices at work in your research so that they too can aspire to
do the best quantitative analyses possible.

References

Al-Hoorie, A. H., & Vitta, J. (2018). The seven sins of L2 research: A review of 30 journals’ statis-
tical quality and their CiteScore, SLR, SNIP, JCR impact factors. Language Teaching Research.
doi:10.1177/1362168818767191
Research, 3, 12–28.
language studies. TESOL Quarterly, 26, 629–664.
Center for Open Science. (2018). Registered reports: Peer review before results are known to align sci-
cientic values and practices [Web page]. Retrieved from https://cos.io/rr
Cumming, G. (2008). Replication and p intervals: P values predict the future only vaguely, but confi-
dence intervals do much better. Perspectives on Psychological Science, 3, 286–300.
Cumming, G. (2012). Understanding the new statistics: Effect sizes, confidence intervals, and meta-


Shawn Loewen and Aline Godfroid


