*Studies in Second Language Acquisition*, 2017, **39**, 579–592.
doi:10.1017/S0272263116000231

# MULTIPLE REGRESSION AS A FLEXIBLE ALTERNATIVE TO ANOVA IN L2 RESEARCH

Luke Plonsky
*Georgetown University*

Frederick L. Oswald
*Rice University*

Second language (L2) research relies heavily and increasingly on ANOVA (analysis of variance)-based results as a means to advance theory and practice. This fact alone should merit some reflection on the utility and value of ANOVA. It is possible that we could use this procedure more appropriately and, as argued here, other analyses such as multiple regression may prove to be more illuminating in certain research contexts. We begin this article with an overview of problems associated with ANOVA; some of them are inherent to the procedure, and others are tied to the way it is applied in L2 research. We then present three rationales for when researchers might turn to multiple regression in place of ANOVA. Output from ANOVA and multiple regression analyses based on published and mock-up studies are used to illustrate major points.

Analysis of variance (ANOVA)[1] is the most frequently used statistical test in quantitative second language (L2) research. A recent methodological synthesis published in *Language Learning* and *Studies in Second Language Acquisition*, for instance, reported that 56% of the 606 quantitative studies in the sample included one or more ANOVAs (Plonsky, 2013) (followed by the *t*-test, which is the two-group analog of ANOVA [43%], then correlation

Correspondence concerning this article should be addressed to Luke Plonsky, Georgetown University, Department of Linguistics, Washington, D.C., 20057. E-mail: luke.plonsky@georgetown.edu

 

[31%], chi-square [19%], and regression [15%]). Moreover, our reliance on ANOVA appears to be growing even stronger in recent years, outpacing increases in other analyses during the same period (Plonsky, 2014).

There is nothing inherently wrong with ANOVA; this procedure is often applied profitably in L2 research and yields very useful information. However, in order for ANOVA to be most informative, a number of assumptions and conditions, both statistical and conceptual in nature, must be met. Unfortunately, as we describe in the following text, these conditions often go unmet or get overlooked entirely. In this article, we review ANOVA in terms of statistical assumptions, statistical power, and data transparency; these issues then contribute to the focal point of the analytic appropriateness of ANOVA for L2 research questions.

## STATISTICAL ASSUMPTIONS

The first and perhaps most obvious prerequisite for ANOVA is the set of statistical assumptions associated with this and many other parametric procedures (e.g., independent observations, normality, and homogeneity of variance). Although simulations show that ANOVA results may not always be sensitive to such assumptions (e.g., Wells & Hintze, 2007), Plonsky's (2013) methodological review of research published 1990–2010 in *Language Learning* and *Studies in Second Language Acquisition* found that researchers reported whether or not assumptions have been checked and met in only about 17% of the sample. Even when assumptions are examined, however, the tests of these assumptions may be statistically underpowered and therefore inaccurate (Field, 2013). It may also be the case that many studies include statistical outliers. When one works with small samples that are typical of L2 research (see Crookes, 1991; Plonsky, Egbert, & LaFlair, 2015), it can be very difficult to tell the difference between an outlier and a reasonable fit of a normal distribution to the data. In short, we need to test the statistical assumptions of ANOVA and not run ANOVAs blindly. As we test these assumptions, we need to have a better understanding about the conditions under which violating ANOVA assumptions will lead to distorted inferences.

## STATISTICAL POWER

We also need enough statistical power to test the aforementioned assumptions, as well as obtain reasonably precise research results. Statistical power is the probability of observing a statistically significant relationship in the data, given that a nonnull relationship exists in the hypothetical (infinite) population that is represented by the sample. Statistical power is a function of three factors: the size of our samples

https://doi.org/10.1017/S0272263116000231 Published online by Cambridge University Press

(e.g., larger *N* raises statistical power), the size of the underlying effect to be detected (e.g., smaller mean differences require more statistical power to be detected), and the alpha cutoff for deciding that a statistical effect exists (smaller alphas require more power to achieve statistical significance). Quantitative L2 research, as reported in Plonsky's (2013) review, currently suffers from a "power problem" resulting from a combination of (a) very small samples (median $n$ = 19 per group), (b) a high rate of null hypothesis significance testing (median number of tests per study = 18), (c) a very low rate of multivariate analyses that can preserve statistical power when the goal is broad and exploratory (e.g., detect an overall effect in multivariate data) or narrow and specified by theory (e.g., detect a specific pattern in multivariate data), (d) effects that are not generally very large (see Plonsky & Oswald, 2014), and (e) an almost complete absence of power analyses prior to the research being conducted, which would generally encourage L2 researchers to collect more data.

These power-related issues and practices, both individually and in concert, pose a serious threat to the internal validity of our research. L2 researchers are conducting theoretically interesting and well-designed studies, and yet if they do not collect enough data, then statistical results will resist the detection of important phenomena. Such results may never get published. With very small samples, a very large effect size is required to reject the null hypothesis; sometimes the effect is so large that one knows even before running the study that it will be very likely to be observed. When such effects are observed, they are (a) more likely to be overstated and (b) more likely to be published. Taken together, these conditions lead to the phenomenon of *publication bias*. But note that even in cases when L2 researchers using ANOVA are able to obtain a large sample and an acceptable level of statistical power to avoid Type II errors, most analyses are still tied in practice to the generally flawed historical paradigm of statistical significance testing. A useful alternative or supplement would be to report and interpret effect sizes and associated confidence intervals (for recent overviews on this issue, see Cumming, 2012; Norris, 2015; Plonsky, 2015b). A Bayesian approach to ANOVA can also incorporate statistical and practical significance (see Mackey & Ross, 2015; Morey, 2015).

## DATA TRANSPARENCY

A third requirement for appropriate use of ANOVA is a thorough summary of the data and, whenever possible, making raw datasets available as well. Transparency in written reports is essential, not only for the readers of primary studies who seek to understand and possibly verify analyses, but also to inform future research in the subdomain, along with future meta-analyses (see, e.g., the guidelines of *Language Learning* for reporting quantitative research: Norris, Plonsky, Ross, & Schoonen, 2015).

https://doi.org/10.1017/S0272263116000231 Published online by Cambridge University Press

A common and serious lack of transparency in L2 research lies in failing to report the standard deviations (*SD*s) associated with the group means being compared (Plonsky & Gass, 2011). Statistically, this prevents the reader from understanding whether the ANOVA assumption of equal within-group variances (and therefore *SD*s) is met (per our previous point). Not reporting *SD*s can also prevent the calculation of standardized mean differences (Cohen's *d*). Sometimes *d* values can be estimated from the *F* test of an ANOVA, but *F* values are often omitted; also, in some situations it makes more sense to calculate *d* values relative to the *SD* of the control group, not the pooled *SD*. Thus, failing to report *SD*s prevents *d* values from being calculated, and this often prevents studies from being included in meta-analyses (Larson-Hall & Plonsky, 2015; Norris & Ortega, 2006) and from undergoing thorough comparisons with findings in subsequent replication studies.

Turning from statistical to conceptual reasons for reporting *SD*s, it helps to know the extent to which individuals within different groups may have responded similarly or differently from one another compared with the group mean. For instance, an intervention group may have promoted everyone to a high mastery level of a target structure, leading to both a higher mean and a reduced *SD* relative to the control group; or the intervention may have increased the mean as well as the *SD* (i.e., some in this group may have improved much more than others).

## ANALYTICAL APPROPRIATENESS

The most critical condition for ANOVA is its fit to the research questions and related data on hand. This condition may appear to be the most obvious and easily satisfied: ANOVA should be used, for example, when a study includes a single independent variable (IV) that is categorical (e.g., experimental condition), and the means of a single dependent variable (DV) across these categories are being compared (see worked examples in Plonsky, in press; Plonsky, 2015a). However, L2 research can fall prey to the same "ANOVA mindset syndrome" (MacCallum, 1998) that psychology has endured, where researchers focus squarely on ANOVA and develop their research questions and data analysis around it, rather than focusing on their research questions and turning to appropriate conceptual, design, and analysis possibilities (ANOVA and otherwise).

On this latter point, it is unfortunate that L2 researchers generally embrace only a very small set of analytic techniques with any regularity outside the realm of ANOVA. Still more concerning, ANOVA is often forced onto the data: for example, dividing scores on a continuously scored measure of motivation or working memory into artificial groups as the IV (e.g., taking a median split on scores to create two groups),

https://doi.org/10.1017/S0272263116000231 Published online by Cambridge University Press

then statistically comparing the means between these artificial groups on a DV such as L2 development or proficiency. When analyses are based on artificial groups like this, any *p*-value, eta-squared, or other statistical result based on an ANOVA can be justifiably questioned or even dismissed out of hand.

Taking a continuous variable and artificially dividing it into two or more groups is a serious mistake, because you lose all the underlying continuous information for no good reason (Cohen, 1983). An analogous situation would be watching the news, where the weather forecaster only tells you that temperatures are either "cooler" or "warmer" with respect to the historical median temperature, without giving you any information on the exact temperatures. Rather than create artificial groups, a more appropriate statistical analysis would likely be based on a correlation or regression analysis of continuous variables.

A study recently published in a prominent L2 journal was interested in the relationship between continuous measures of working memory and reading comprehension. Rather than correlate them directly, the author divided participants into three groups based on composite working memory scores (i.e., low, medium, and high scoring groups). A series of more than 30 ANOVAs and *t*-tests was then conducted to determine how the working memory groups and text types (another IV) led to differences in average reading comprehension. Among other results, the analyses revealed an eta$^2$ value of .10 ($p < .05$). Critically, this result does not provide us with direct information on the nature of the relationship between working memory and reading comprehension. It only tells us that 10% of the variance in reading scores can be attributed to group membership across three artificially formed groups or levels of working memory. The purpose here is not to highlight the weakness of this particular study, but rather to illustrate the way researchers mistakenly mold their data to fit the ANOVA approach.

The field's use of ANOVA also tends to outpace its utility when more than one IV is included in the design, as shown in the previous example (working memory and text type) and as is often the case given the inherently multivariate nature of L2 learning and use. Nevertheless, it is exceedingly common to find studies that ignore relationships between the independent variables. Plonsky's (2013) review of analytical practices, for instance, found a median of 18 such tests per study. By relying so heavily on a series of univariate ANOVAs, researchers are ignoring the correlated (partially redundant) nature of the relationships of interest; they are weakening what already is limited statistical power, and they are increasing their chances of committing Type I errors.

Imagine a study in which the researcher was interested in understanding L2 vocabulary knowledge in relation to three theoretically motivated variables: first language (L1), length of study, and motivation. Let's assume the sample includes 90 participants representing three

https://doi.org/10.1017/S0272263116000231 Published online by Cambridge University Press

different L1s equally: English ($n$ = 30), Vietnamese ($n$ = 30), and Spanish ($n$ = 30). The two other variables, length of study and motivation, are represented by continuous measures. The descriptive statistics for each variable in this mock-up study are presented in Tables 1 and 2. For anyone interested in rerunning these analyses, the dataset will be made available upon request as well as on the IRIS Database (see Marsden, Mackey, & Plonsky, 2016).

A conventional approach in this situation would be to run three tests: an ANOVA to compare vocabulary knowledge across the three L1 groups, and two correlations to measure the relationship between vocabulary knowledge with both length of study and motivation. (Some researchers may also mistakenly divide up the sample further into subgroups, based on participants' length of study and/or motivation scores, to allow for additional ANOVAs. This practice, described in the preceding text, involves converting continuous variables into categorical ones, resulting in an unnecessary loss of data.) These tests yield the following results for the relationship between our DV, vocabulary knowledge, and (a) L1 ($F$ = 3.04, $p$ = .05, eta$^2$ = .07), (b) length of study ($r$ = .87, $p$ < .001), and (c) motivation ($r$ = .54, $p$ < .001).

Many researchers at this point would also conduct post hoc contrasts for the L1 backgrounds and then conclude their analyses, satisfied to report several statistically significant results. However, doing so would ignore the potential relationships between the IVs. Reporting three significant findings based on IVs that are all highly correlated amounts to testing essentially the same relationship three times. And even when IVs are correlated but not as highly, the redundancy should still be statistically accounted for. In this example, it is worth investigating whether length of study and motivation scores are correlated. One or both of these variables might also differ across L1 groups. If observed, such relationships imply that it is not appropriate to analyze the four variables in this study in a bivariate fashion. Rather, a more comprehensive approach such as multiple regression (demonstrated in the following text) would likely be more appropriate and informative.

We should recognize at this point that factorial ANOVA and analysis of covariance (ANCOVA) can examine interactions between IVs. In the mock-up study, we might see an interaction between L1 and proficiency

**Table 1.** Vocabulary knowledge scores across L1 groups

| | *M* (*SD*) | 95% CIs |
|---|---|---|
| English | 3.07 (1.60) | [2.47, 3.66] |
| Vietnamese | 4.13 (1.66) | [3.52, 4.75] |
| Spanish | 3.63 (1.73) | [2.99, 4.28] |
| Total | 3.61 (1.70) | [3.25, 3.97] |

https://doi.org/10.1017/S0272263116000231 Published online by Cambridge University Press

**Table 2.** Descriptive statistics for length of study and motivation scores

| | *M* (*SD*) | 95% CIs |
|---|---|---|
| Length of Study | 8.28 (4.94) | [7.24, 9.31] |
| Motivation | 4.30 (1.63) | [3.96, 4.64] |

level (assuming participants had also been tested for the latter). More often than not, however, this approach leads to a series of statistical results that require additional statistical power, and that can be difficult for both authors and readers to interpret beyond the main effects. The other main weakness in factorial ANOVA is that most L2 researchers who use it maintain a sole interest in the presence or absence of statistically significant mean differences, despite the readily available effect size index (generally eta$^2$, or partial eta$^2$ in the case of multiple IVs; see Norouzian & Plonsky, in press) that indicates variance accounted for in the DV as a function of group membership on one or more IVs.

To summarize the argument made thus far, quantitative L2 research relies very heavily on an analytical approach that in our view is often not appropriate to the data and/or that is not utilized or reported on appropriately. By adhering to ANOVA, the potential of our empirical efforts to inform and advance L2 theory and practice is obstructed. More specifically, rather than examining and explaining the variance in DVs as a function of IVs, we concentrate almost exclusively on mean differences.

One path toward more appropriate data analyses involves a recognition of ANOVA, regression, and most other statistics used by L2 researchers as part of a larger statistical framework referred to as the general linear model (GLM). Nearly half a century ago, Jacob Cohen (1968) presented to the field of psychology an argument with a very similar message to that of the current article: that multiple regression, as a parent procedure of ANOVA, provides the same information as ANOVA, as well as a number of improvements. As described and demonstrated quite clearly by Skidmore and Thompson (2010), analyses employed by psychologists in the decades that followed Cohen's now-classic paper were characterized by a marked shift away from ANOVA/ANCOVA and toward multiple regression, which can incorporate continuous variables and also yield interpretable results. L2 researchers would do well to consider a similar trajectory. In the remainder of the article, we outline three reasons why multiple regression can and should be applied in place of ANOVA in many instances as a means to produce analyses and results that have greater potential to inform L2 theory and practice.

https://doi.org/10.1017/S0272263116000231 Published online by Cambridge University Press

## THREE REASONS TO USE MULTIPLE REGRESSION

### The Multivariate Nature of L2 Research

The first and perhaps most compelling reason to turn to multiple regression in place of ANOVA is conceptual (rather than statistical) in nature: the constructs and processes central in L2 research—learning, teaching, use, and assessment—are almost always multivariate in nature, and L2 researchers increasingly are measuring more of this multivariate space. In order to understand multivariate data, it is often necessary to employ analytical procedures that incorporate the simultaneous relationships predicted by theory and represented in the data (Brown, 2015). Not to do so is counterintuitive and unnecessarily narrow and naïve.

Consider the last study you read or one that you are currently working on. Are two or more IVs involved? If so, might there be relationships (correlations) between them? Go back to our mock-up study, where learners' vocabulary knowledge was examined as a function of their L1, length of study, and motivation. Nothing prevents a researcher from addressing the relationship between each of these IVs (or predictors) and the DV (vocabulary knowledge) separately, as shown in the example analyses. However, if the IVs are correlated, as is often the case, this approach will lead us to overstate our understanding of vocabulary knowledge. Specifically, the effect size indices from those analyses (eta$^2$ = .07 for L1, $r^2$ = .76 for length of study, and $r^2$ = .29 for motivation) do not and cannot account for the shared variance between the IVs. The overlapping contributions of the predictor variables is also evident in the three effect size estimates, which add up to 112% because the predictors are related (e.g., length of study and motivation correlate at $r$ = .49).

It is precisely with these reasons in mind that sociolinguists, working in the variationist tradition, often turn to regression models to explain variable structures such as subject expression in Spanish. Rather than examine contextual (e.g., age and socioeconomic status) and linguistic (e.g., animacy, tense, and mood) predictors of variable structures in isolation, the use of (logistic) regression allows variationists to examine variants as a function of such predictors (e.g., Geeslin & Guijarro-Fuentes, 2008).

In contrast to a one-variable-at-a-time approach, multiple regression results based on the same data (with dummy coding of the categorical variable, L1) allow us to account for relationships between predictors and thereby estimate their relative contributions to variance in the dependent variable. Standard multiple regression based on the full set of possible predictors yields an overall $R^2$ value of .80 (Table 3).

https://doi.org/10.1017/S0272263116000231 Published online by Cambridge University Press

**Table 3.** Regression results for predictors of L2 vocabulary knowledge

| Variable | *B* | *SE B* | $\beta$ | *t* | $\Delta R^2$ |
|---|---|---|---|---|---|
| Length | .27 | .02 | .79 | 13.81* | .440 |
| L1 Spanish | .74 | .20 | .21 | 3.70* | .031 |
| Motivation | .14 | .06 | .13 | 2.20* | .011 |
| L1 Vietnamese | .26 | .21 | .07 | 1.23 | .003 |

*Note*. Adjusted $R^2$ = .80 ($N$ = 90, *$p$ < .05); L1 English = reference group.

As we might expect based on the bivariate correlations from the preceding text, the largest standardized beta coefficient ($\beta$) results from length of study: .79. We can interpret this value as indicating that for every year of unit increase in the predictor variable (i.e., every one-unit increase in this variable), we can expect an increase in vocabulary knowledge of .79 *SD* units, controlling for other variables in the model. Length of study also yields the largest $\Delta R^2$ value (.440), which expresses the variance in vocabulary scores accounted for by the predictor if added last to the model. Both $\beta$ and $\Delta R^2$ can be helpful in interpreting regression results, though their uses and interpretations differ. Whereas $\beta$ is useful in making predictions for values in the criterion (dependent) variable, after accounting for other variables in the equation, $\Delta R^2$ allows the researcher to examine the unique contribution of the variable that it is associated with. L1 Spanish, but not L1 Vietnamese, is also a significant and positive predictor, indicating that L1s are differentially related to vocabulary knowledge.[2] And although motivation was strongly correlated with vocabulary knowledge ($r$ = .54), its predictive power was much weaker when accounting for the variance it shares with the other predictor variables. Consequently, the $\Delta R^2$ associated with motivation is .011, which is quite small, indicating that this variable only explains an additional 1% of variance in the DV.

If we are interested in arriving at a statistical model that best represents the theoretical relationships being examined and our data, a regression approach is often both more informative and more appropriate than a series of univariate or bivariate analyses.

## Variance Matters

As mentioned previously, the convention of L2 researchers to rely on means-based comparisons fails us in at least two ways. First, the default status of ANOVA often leads researchers to artificially and arbitrarily reduce continuously measured variables into categorical ones. By doing so, we sacrifice precious and meaningful variance for what may appear

https://doi.org/10.1017/S0272263116000231 Published online by Cambridge University Press

to be a more straightforward analytical approach. Regression analyses including simple bivariate correlations, by contrast, allow for the variance in scores to be preserved.

Consider the published example study from the preceding text in which working memory scores were used to place participants into groups to then compare them on a measure of reading comprehension using ANOVA. Because the groups were formed arbitrarily, resulting in a loss of variance in the IV, the analysis produced a relatively crude measure of the relationship between working memory and reading comprehension. A simple correlation would have been both simpler and superior in its ability to characterize and quantify the relationship between these two variables.

Our reliance on mean differences also leads us to ignore the variance accounted for in dependent or criterion variables. Although eta$^2$ and partial eta$^2$ are now regularly reported along with ANOVAs, these values are rarely interpreted in terms of variance and are most often either ignored or generically labeled as referring to a small, medium, or large effect. L2 theory can only proceed so far on a diet of mean differences; advancing theory related to key constructs such as vocabulary knowledge, accentedness, or instructional effects requires an understanding of their variance as reflected by variables such as individual differences and length of exposure/treatment. Uncovering mean differences can be useful and interesting, but seeking to identify and estimate other sources of variance is a worthwhile goal. A one-way ANOVA is focused on between-group variance, which is expressed as a mean difference; within-group variance is viewed as error variance in the model, and yet this variance can still reflect reliable individual differences in the DV.

## Regression Can Do Everything ANOVA Can Do, and More

In both conceptual and statistical terms, one-way ANOVA is analogous to a specific type of multiple regression: if there are $k$ groups, then there are $k - 1$ predictors coded 0/1 (0 = does not belong to the group, 1 = belongs to the group; i.e., "dummy coding"). Within this multiple regression framework, the ANOVA information is provided: the output yields the same effect size index ($R^2$, equivalent to eta$^2$), and the same $p$ value associated with that effect, indicating whether group membership on a categorical IV explains variance in the DV (see examples from sample studies in the preceding text). However, regression provides a more flexible framework than one-way ANOVA: regression can incorporate additional continuous and categorical IVs (usually called predictor variables in the context of regression), and the individual and combined beta and $R^2$ values for multiple regression are rather

https://doi.org/10.1017/S0272263116000231 Published online by Cambridge University Press

straightforward to interpret. For example, Egbert and Plonsky (2015) were interested in exploring conference abstract ratings as a function of linguistic and stylistic features such as length (in words), use of first-person pronouns, and the presence of four discourse "moves": introduction, methods, results, and discussion. Each of these variables could have been examined separately in relation to abstract ratings using a series of ANOVAs and other tests. However, because of the potential for correlations between predictors (e.g., length with moves), and as a means to embrace a more integrated analytic approach and set of results, the authors employed multiple regression instead. Results indicated that a total of 31% of the variance in abstract scores could be accounted for by a set of six linguistic and stylistic features of abstracts: more words ($R^2$ = .14), citations (.07), the presence of a results section (.04), more nouns (.03), no errors (.02), and fewer first-person pronouns (.01).

## CONCLUSION

This article has argued against the use of ANOVA as the default analytical approach in quantitative L2 research. We have also proposed multiple regression as an alternative that can help L2 researchers both (a) overcome challenges inherent to ANOVA and (b) make fuller and more flexible use of the information contained in their data (Perkins & Newman, 2014).

Given the state of quantitative and methodological literacy in the field (Gonulal, 2016; Loewen et al., 2014), some scholars may interpret this article as a call for greater statistical sophistication. But it is not. Multiple regression has been embraced for decades in nearly all fields of quantitative research. And to be clear, we are not arguing for the blind proliferation of statistical analyses that do not increase our knowledge, in the end. We are all in favor of the statistical less-is-more approach embodied by the American Psychological Association (e.g., Wilkinson & Task Force on Statistical Inference, 1999) and by the recent author guidelines of *Language Learning* (Norris et al., 2015).

We also understand that if multiple regression is to make its way into our regular repertoire of quantitative techniques, we need to give serious consideration to the challenges that it might introduce. Brown (2015) reminds us, for example, that more sophisticated procedures often require more and/or more stringent assumptions, larger samples, and a greater role for the researcher in terms of interpreting results (see Nathans, Oswald, & Nimon, 2012). With these concerns in mind, Jeon (2015) and Larson-Hall (2015) both provide accessible, step-by-step guides to conducting regression analyses. Another conceptual challenge stems from the field's traditional conventions and practices.

https://doi.org/10.1017/S0272263116000231 Published online by Cambridge University Press

Moving forward, we need to consider the value of understanding variance, not just mean differences. This point can apply even in the context of quasiexperimental studies that compare posttest scores between groups. For instance, as illustrated in Plonsky and Ziegler (2016), a mean difference effect of $d \approx .50$ can be visualized using basic tools (e.g., http://rpsychologist.com/d3/cohend/) or expressed as an alternate effect size index such as $U_3$ to indicate the degree of nonoverlap between groups (≈70%; see Lipsey et al., 2012).

Finally, in addition to multiple regression, there are a number of other statistical procedures, such as mixed-effects models (Cunnings & Finlayson, 2015; Gries, 2015) and structural equation modeling (e.g., Hancock & Schoonen, 2015; Schoonen, 2015), many of which also belong to the GLM, that are rarely used but that could be applied fruitfully and more often in L2 research. The focus here has been on multiple regression as a relatively straightforward alternative to ANOVA, by far the most commonly applied statistic and research mind-set in the field.

*Received 11 November 2016*
*Accepted 11 May 2016*
*Final Version Received 18 May 2016*

## NOTES

1. Our use of "ANOVA" in this article refers to both one-way analysis of variance as well as to factorial models unless otherwise specified. Also, when referring to more than one analysis of variance, we use "ANOVAs"; otherwise the singular form ("analysis") can be assumed.

2. The third L1 in this example, English, was treated as a reference group to which the other two groups were compared. In such cases, the descriptive statistics (Table 1) can be helpful in interpreting group scores relative to one another. As pointed out by an anonymous reviewer, ANOVA and regression analyses carried out in R allow the analyst to move relatively seamlessly between the two, thus facilitating interpretations of results involving categorical predictor variables.

## REFERENCES

Brown, J. D. (2015). Why bother learning advanced quantitative methods in L2 research. In L. Plonsky (Ed.), *Advancing quantitative methods in second language research* (pp. 9–20). New York, NY: Routledge.

Cohen, J. (1968). Multiple regression as a general data-analytic system. *Psychological Bulletin*, *70*, 426–443.

Cohen, J. (1983). The cost of dichotomization. *Applied Psychological Measurement*, *7*, 249–253.

Crookes, G. (1991). Power, effect size, and second language research. Another researcher comments. *TESOL Quarterly*, *25*, 762–765.

Cumming, G. (2012). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. New York, NY: Routledge.

Cunnings, I., & Finlayson, I. (2015). Mixed effects modeling and longitudinal data analysis. In L. Plonsky (Ed.), *Advancing quantitative methods in second language research* (pp. 159–181). New York, NY: Routledge.

Egbert, J., & Plonsky, L. (2015). Success in the abstract: Exploring linguistic and stylistic predictors of conference abstract ratings. *Corpora*, *10*, 291–313.
Field, A. (2013). *Discovering statistics using SPSS* (4th ed.). Thousand Oaks, CA: Sage.
Geeslin, K. L., & Guijarro-Fuentes, P. (2008). Variation in contemporary Spanish: Linguistic predictors of *estar* in four cases of language contact. *Bilingualism: Language and Cognition*, *11*, 365–380.
Gonulal, T. (2016). Statistical literacy among second language acquisition graduate students. Unpublished doctoral dissertation, Michigan State University, East Lansing, MI.
Gries, S. Th. (2015). The most underused statistical method in corpus linguistics: Multi-level (and mixed-effects) models. *Corpora*, *10*, 95–125.
Hancock, G. R., & Schoonen, R. (2015). Structural equation modeling: Possibilities for language learning researchers. *Language Learning*, *65*(Supp. 1), 160–184.
Jeon, E. H. (2015). Multiple regression. In L. Plonsky (Ed.), *Advancing quantitative methods in second language research* (pp. 131–158). New York, NY: Routledge.
Larson-Hall, J. (2015). *A guide to doing statistics in second language research using SPSS and R* (2nd ed.). New York, NY: Routledge.
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.
Lipsey, M. W., Puzio, K., Yun, C., Hebert, M. A., Steinka-Fry, K., Cole, M. W., et al. (2012). *Translating the statistical representation of the effects of education interventions into more readily interpretable forms* (NCSER 2013-3000). Washington, DC: National Center for Special Education Research, Institute of Education Sciences, U.S. Department of Education.
Loewen, S., Lavolette, E., Spino, L. A., Papi, M., Schmidtke, J., Sterling, S., et al. (2014). Statistical literacy among applied linguists and second language acquisition researchers. *TESOL Quarterly*, *48*, 360–388.
MacCallum, R. (1998). Commentary on quantitative methods in I/O research. *Industrial-Organizational Psychologist*, *35*, 19–30.
Mackey, B., & Ross, S. J. (2015). Bayesian informative hypothesis testing. In L. Plonsky (Ed.), *Advancing quantitative methods in second language research* (pp. 329–345). New York, NY: Routledge.
Marsden, E., Mackey, A., & Plonsky, L. (2016). Breadth and depth: The IRIS repository. In A. Mackey & E. Marsden (Eds.), *Advancing methodology and practice: The IRIS repository of instruments for research into second languages* (pp. 1–21). New York, NY: Routledge.
Morey, R. (2015). Multiple comparisons with BayesFactor, part 1 [web log post]. Retrieved from http://bayesfactor.blogspot.co.uk/2015/01/multiple-comparisons-with-bayesfactor-1.html.
Nathans, L. L., Oswald, F. L., & Nimon, K. (2012). Interpreting multiple linear regression: A guidebook of variable importance. *Practical Assessment, Research, & Evaluation*, *17*, 1–19.
Norouzian, R., & Plonsky, L. (in press). Eta- and partial eta-squared in L2 research: A cautionary review and guide to more appropriate usage. *Second Language Research*.
Norris, J. M. (2015). Statistical significance testing in second language research: Basic problems and suggestions for reform. *Language Learning*, *65*(Supp. 1), 97–126.
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. 3–50). Philadelphia: John Benjamins.
Norris, J. M., Plonsky, L., Ross, S. J., & Schoonen, R. (2015). Guidelines for reporting quantitative methods and results in primary research. *Language Learning*, *65*, 470–476.
Perkins, K., & Newman, I. (2014). How multiple regression models can be written to better reflect the complexities of first and second language acquisition research: An attempt to limit Type VI error. *Multiple Linear Regression Viewpoints*, *40*, 41–51.
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*, 655–687.
Plonsky, L. (2014). Study quality in quantitative L2 research (1990–2010): A methodological synthesis and call for reform. *Modern Language Journal*, *98*, 450–470.

Plonsky, L. (2015a). Quantitative considerations for improving replicability in CALL and applied linguistics. *CALICO Journal*, *32*, 232–244.
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. (in press). Quantitative research methods. In S. Loewen & M. Sato (Eds.), *The Routledge handbook of instructed second language acquisition*. New York, NY: Routledge.
Plonsky, L., & Gass, S. (2011). Quantitative research methods, study quality, and outcomes: The case of interaction research. *Language Learning*, *61*, 325–366.
Plonsky, L., & Oswald, F. L. (2014). How big is "big"? Interpreting effect sizes in L2 research. *Language Learning*, *64*, 878–912.
Plonsky, L., & Ziegler, N. (2016). The CALL-SLA interface: Insights from a second-order synthesis. *Language Learning & Technology*, *20*, 17–37.
Plonsky, L., Egbert, J., & LaFlair, G. T. (2015). Bootstrapping in applied linguistics: Assessing its potential using shared data. *Applied Linguistics*, *36*, 591–610.
Schoonen, R. (2015). Structural equation modeling in L2 research. In L. Plonsky (Ed.), *Advancing quantitative methods in second language research* (pp. 213–242). New York, NY: Routledge.
Skidmore, S. T., & Thompson, B. (2010). Statistical techniques used in published articles: A historical review of reviews. *Educational and Psychological Measurement*, *70*, 777–795.
Wells, C. S., & Hintze, J. M. (2007). Dealing with assumptions underlying statistical tests. *Psychology in the Schools*, *44*, 495–502.
Wilkinson, L., & Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, *54*, 594–604.