

3

STATISTICAL POWER, *P* VALUES, DESCRIPTIVE STATISTICS, AND EFFECT SIZES

A “BACK-TO-BASICS” APPROACH TO ADVANCING QUANTITATIVE METHODS IN L2 RESEARCH

Luke Plonsky

Introduction

Methodologically speaking, a great deal of quantitative L2 research has been misguided. All too often we have been asking the wrong questions of our data. Consequently, many of the answers we have derived have been, at best, weak in their ability to inform theory and practice and, at worst, wrong or misleading. This chapter seeks to reorient the field toward more appropriate kinds of questions and analytical approaches. More specifically, I argue here against the field's flawed use and interpretation of statistical significance and, instead, in favor of more thorough consideration of descriptive statistics including effect sizes and confidence intervals (CIs). The approach I advocate in this chapter is not only more basic, statistically speaking, and more computationally straightforward, but it is also inherently more informative and more accurate when compared to the most fundamental and commonly used analyses such as *t* tests, ANOVAs, and correlations.

I begin the chapter with a model that describes quantitative L2 research as currently practiced, pointing out major flaws in our approach. I then review major weaknesses of relying on statistical significance (*p* values), particularly in the case of tests comparing means (*t* tests, ANOVAs) and correlations. I follow this discussion with a brief introduction to the notion of statistical power, followed by guides to calculating and using effect sizes and other descriptive statistics including CIs. I conclude with a revised/proposed model of what quantitative L2 research might look like if we were to embrace this approach. Points made throughout the discussion are illustrated with data-based examples, many of which can be replicated using the practice data set that accompanies this chapter (<http://oak.ucc.nau.edu/ldp3/AQMSLR.html>). Unlike much of the remainder of this book, the statistical issues in this chapter are very simple. Nevertheless, these ideas largely go against what is often taught in introductory research methods courses and certainly what is found in most L2 journals.

Before beginning the main discussion, I also want to emphasize that the concepts and procedures in this chapter, though far from mainstream L2 research practice, are central to a set of methodological reforms currently gaining traction in the field. Among other issues, this movement has sought to (a) encourage replication research (Porte, 2012), (b) promote a synthetic ethic in primary as well as secondary research (e.g., Norris & Ortega, 2000, 2006; Oswald & Plonsky, 2010; Plonsky & Oswald, Chapter 6 in this volume), (c) critically reflect on and examine methodological practices and self-efficacy (e.g., Larson-Hall & Plonsky, 2015; Loewen et al., 2014; Plonsky, 2013, 2014), and (d) introduce novel analytical techniques (e.g., Cunnings, 2012; Larson-Hall & Herrington, 2010; LaFlair, Egbert, & Plonsky, Chapter 4 in this volume; Plonsky, Egbert, & LaFlair, in press). Taking yet another step back, it is also worth noting that, although many of the concepts and techniques embodied by this movement and discussed in this chapter may be unfamiliar to L2 researchers, they have been recognized for decades as the preferred means to conducting basic quantitative research among methodologists in other social sciences such as psychology and education.

The Flawed Notion of Statistical Significance

To begin this discussion on the flaws of statistical significance, let's first consider the pivotal role of p values. Figure 3.1 presents a descriptive account of the path by which most quantitative L2 research attempts to advance the field. Researchers begin by conducting a study on the effect of A on B or the relationship between X and Y. (Note: Most studies are already flawed at this point in that their research questions elicit only yes/no answers such as "Is there a difference . . . ?" or "Is there a relationship between . . . ?". A much more informative approach is to pose more open-ended research questions that are inherently more informative and that better represent the continuous data being collected, such as "To what extent . . . ?".

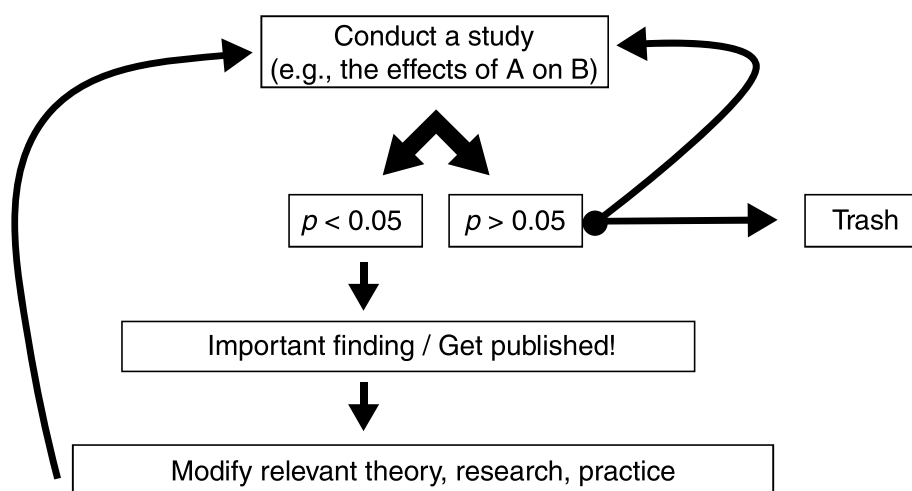


FIGURE 3.1 A descriptive model of quantitative L2 research

Once the data are collected and analyzed using, for example, a t -test or Pearson correlation, most researchers will take special note of the p value associated with the results of those tests. As depicted in Figure 3.1, if on one hand the p value is larger than .05, the difference between groups or the correlation is often considered uninteresting and is discarded, and another study might be run to attempt to achieve a statistically significant result. On the other hand, if the t -test or correlation yields a statistically significant result (i.e., $<.05$), it is considered important and is much more likely to get published and to consequently have an impact on L2 theory, future research, and practice.

In this model, which is, again, the dominant approach in quantitative L2 research, researcher perception and dissemination of study results both hinge critically on our adherence to null hypothesis significance testing (NHST). As I describe in the remainder of this section, this approach is deeply flawed on many accounts, both conceptually and statistically. I focus here, though, on three main arguments: (a) NHST is unreliable, (b) NHST is crude and uninformative, and (c) NHST is arbitrary. Among the many other, more comprehensive accounts of the inherent flaws in NHST, I recommend Kline (2013, Chapter 3), Norris (in press), and Cumming (2012, Chapter 2).

NHST Is Unreliable

The first major flaw of NHST is that it is unreliable. More specifically, because p values vary as a function of sample size, any correlation or difference in mean scores can reach statistical significance, given a large enough sample. Consider the (fabricated) data from three studies in Tables 3.1–3.3 each of which, let’s say, is interested in comparing the effects of traditional (Group 1) with experimental (Group 2) approaches to teaching vocabulary. A t -test comparing the means in Study 1 found no difference between the two groups, which each have five participants.

Study 2 collected data from 15 participants in each condition. Although their means and standard deviations were identical to those in Study 1, the p value in

TABLE 3.1 Data and results from Sample Study 1

<i>Study</i>	N_1	N_2	$M_1 (SD_1)$	$M_2 (SD_2)$	p	d
Study 1	5	5	15 (3)	18 (4)	.2265	.85

TABLE 3.2 Data and results from Sample Study 2

<i>Study</i>	N_1	N_2	$M_1 (SD_1)$	$M_2 (SD_2)$	p	d
Study 1	15	15	15 (3)	18 (4)	.0276	.85

TABLE 3.3 Data and results from Sample Study 3

<i>Study</i>	N_1	N_2	$M_1 (SD_1)$	$M_2 (SD_2)$	p	d
Study 1	45	45	15 (3)	18 (4)	.0001	.85

Study 2 was found to be statistically significant. The results of this study, therefore, indicate that there is a real difference between the two conditions.

The samples in Study 3 were even larger: 45 participants in each group. The same means and standard deviations were observed for the two conditions again, but this time the p value is even smaller: .0001.

Although the only difference across studies was in the sample size, in a traditional NHST framework, we would likely interpret these studies as showing inconsistent support for the experimental intervention. If we focus on the descriptive statistics, including the Cohen's d effect size, however, we see that all three studies provide the same exact result: a positive and somewhat strong effect ($d = .85$) in favor of the experimental condition. The only difference across studies was in the sample size.

The same inconsistency we observed in the previous example also applies to correlational analyses (and virtually all other analyses based on NHST). A correlation coefficient of .4 based on 30 participants may not be statistically significant. With a sample of 60, however, that same correlation would, in most cases, yield a p value below .05. In both cases, the correlation between the two variables as observed is .4, but the statistical significance is different simply because of their respective N s.

At this point and as a means to help make sense of the potential unreliability of results simply based on different sample sizes, it might be useful to remind ourselves of the definition of p : the likelihood that the observed mean difference (or correlation, etc.) would be observed given a true population difference (or correlation) of 0 (i.e., $d = 0$; $r = 0$). Because neither the mean difference nor the correlation is ever going to be 0, any size mean difference or correlation can reach statistical significance given a large enough sample. Along these lines, over two decades ago, Thompson (1992) reminds us that “[with NHST] . . . tired researchers, having collected data on hundreds of subjects, then conduct a statistical test to evaluate whether there were a lot of subjects, which the researchers already know, because they collected the data and know they are tired.” (p. 434).

NHST Is Crude and Uninformative

The results provided by NHST are not only unreliable, they are extremely uninformative. When we submit our data to a test of statistical significance, we reduce the number of possible outcomes to two. In other words, we reduce continuous

results to a yes/no dichotomy, often overlooking or even ignoring the rich information provided by our descriptive statistics. By doing so, we waste our data and we fail to accurately or informatively advance L2 theory, research, and practice.

To be sure, p values tell us nothing about (a) replicability, (b) theoretical or practical importance, or, perhaps most importantly, (c) magnitude of effects. A p value of greater than .05 does not necessarily indicate that there is no difference between two group means or even that there is a small difference between two group means. Nevertheless, many researchers interpret it that way, falling prey to what Cumming (2012) calls the “slippery slope of nonsignificance” (p. 31). Likewise, very small p values can certainly correspond to small effects.

To illustrate the lack of informational value provided by p values, consider the following examples from published L2 studies. In one study published recently the authors present the results of a t -test comparing the “ideal L2 self” ratings for high- ($M = 4.65$, $SD = 1$) and low-motivation ($M = 4.56$, $SD = 1.1$) learners. The t -test yielded a nonstatistically significant p value, indicating no difference between the two groups. This result, to be expected given the very similar descriptives, was confirmed by a very small eta-squared effect size of .002, which we can understand to mean that group membership (i.e., high vs. low motivation) explains less than 1% of the variance in ideal self ratings. In the same table, the authors present the results of another t -test comparing the same two groups on ought-to self ratings. The mean score was 3.74 ($SD = 1.1$) for the high-motivation group and 3.96 ($SD = 1$) for the low-motivation group. In this case, however, the t -test revealed a statistically significant difference between the groups. Are we then to interpret the difference between groups here to be large or important? The eta-squared value for this contrast was just .01, indicating that group membership could explain 1% of the variance in group means. From a dichotomous NHST perspective, one of these tests reveals an important difference in group means and the other does not. From the perspective of practical significance based on the effect size and other descriptive statistics, it is clear that the two groups are nearly identical. (See results related to Table 4 in Mackey & Sachs, 2012, for a counterexample wherein the authors correctly interpret substantial correlations despite the nonstatistical p values associated with them.)

Consider as well the results in Table 3.4 which were extracted from nine primary studies in Taylor, Stevens, and Asher’s (2006) meta-analysis of the effects of reading strategy instruction. Three distinct patterns of results can be observed in this sample, each of which reveals the crudeness of p . First, although the means being compared in studies A–E were not found to be statistically significant, their effect sizes (Hedges’ g , which expresses mean difference in standard deviation units, similar to Cohen’s d) were substantial—certainly more than the null effect we might interpret based on a nonstatistical p value. These effect sizes were, in fact, almost identical to but slightly larger, actually, than those in

TABLE 3.4 Example results showing the inconsistency of *p* values*

<i>Study</i>	N_1	N_2	<i>Effect size (g)</i>	<i>p</i>
A	12	15	-.555	.152
B	8	8	.556	.259
C	30	29	.492	.060
D	24	21	.553	.066
E	21	22	.472	.123
F	78	80	.481	.003
G	183	61	.530	.000
H	29	14	-.251	.436
I	12	14	-.292	.450

*Results from Taylor et al. (2006)

studies F and G, both of which obtained statistical significance. Second, recall from the previous section that *p* values fluctuate as *N*s increase or decrease. In this particular case, although the effect sizes from A–E and F–G are very similar, the *p* values in the latter group are statistically significant due to their relatively large *N*s. Third, like studies A–E, H and I yielded *p* values larger than .05. In the NHST approach, these results would therefore be equated with no difference between groups. However, not only does the effect size show a nontrivial difference between groups; these differences and that of A run in the opposite direction to what we might expect, showing a substantial advantage for the comparison groups. Bottom line: Not only are *p* values unreliable, but they also fail to provide information about the size or importance of the relationships and effects we are interested in.

NHST Is Arbitrary

Students in introductory research methods courses often ask what is so special about the .05 level of statistical significance. The answer, of course, is nothing—a sentiment Rosnow and Rosenthal (1989) had in mind when they quipped, “surely, God loves the .06 nearly as much as the .05” (p. 1277). Nevertheless, much of the field lives (or least publishes) according to an arbitrary standard for importance.

To summarize the discussion thus far, quantitative L2 research relies very strongly on an analytical approach that is unreliable and arbitrary. Even if NHST-based findings were stable and principled, results based on this approach would still fail to provide us any indication of the kinds of information we are most interested in or that can guide L2 theory and practice. Consequently, unless we are content to attempt to advance our field in this fashion (i.e., based on arbitrary, unreliable, yes/no-only results), we must change our approach (see Norris, *in press*).

Statistical Power

A closely related notion, statistical power is the probability of observing a statistically significant relationship given that the null hypothesis is false (e.g., $d \neq 0$; $r \neq 0$). The more powerful the study, the less likelihood of false negatives. An understanding of power can also be used to answer the very practical and frequent question of “How many participants do I need (to detect statistical significance)?” (That is, assuming we are still interested in statistical significance.)

The conventionally desired level of statistical power in the social sciences is .80 which, when achieved, provides the researcher with an 80% chance of detecting a statistical relationship if present (Cohen, 1992). (Note that the .80 convention for avoiding false negatives is much more liberal than the typical safeguard for avoiding false positives of .05. In the former, we implicitly accept an error rate of 20%; in the latter the accepted error rate is theoretically only 5%.) But how can we determine if .80 power is possible? As with statistical significance, power varies as a function of the effect size and sample size such that, given a larger anticipated effect (e.g., $d \approx 1$), a smaller sample will be able to detect a statistical relationship 80% of the time ($N \approx 35$). Likewise, when a small effect (e.g., $d \approx .2$) is expected based on theoretical predictions and/or previous research, a larger sample ($N \approx 400$) is needed in order to have an 80% chance of finding the effect at the .05 level.¹

A related exercise and consideration might be to estimate the statistical power in previous L2 research, much of which relies necessarily on small samples. Plonsky and Gass (2011) examined this issue by means of a post hoc power analysis for 174 studies in the interactionist tradition of L2 research. Their results show that this subdomain has had, on average, just a 56% chance of obtaining statistically significant results. Likewise, looking at 606 primary studies across many different subdomains of L2 research, Plonsky (2013) found average post hoc power at just .57. These results can be interpreted as indicating that the likelihood of observing expected relationships is, on average, comparable to tossing a coin and hoping for heads.

Evidence of what I refer to as the “power problem” (Plonsky, 2013, p. 678) in L2 research does not stop there. Additional indications include (a) extremely rare use of power analyses in order to inform sampling decisions, (b) generally small samples / high sampling error, (c) heavy reliance on NHST, (d) presence of non-normal distributions and a lack of checking for statistical assumptions, and (e) relatively infrequent use of multivariate statistics that can preserve experiment-wise power (Plonsky, 2013).

One step toward addressing this problem is to determine sample sizes based on a priori power analyses, rather than simply based on convenience or convention. Using free software such as G*Power (Faul, Erdfelder, Lang, & Buchner, 2007) or any number of freely available online calculators designed for this purpose, you can calculate the sample size needed for a given level of statistical power such as

.80. The only information you need to bring to the equation is the anticipated effect size. One source for obtaining this value would be a meta-analysis on a topic closely related to that of the study. In the absence of a relevant meta-analytic effect size, you could also plug into the equation the effect size from one or more studies on a closely related topic.

At this point I should recognize that in some instances it is not possible to collect data from a sample large enough to obtain an ideal level of statistical power. For example, researchers who study learners of less commonly taught languages may find it difficult to obtain large samples. Similarly, funding may not be available to pay as many participants as are needed for adequate power. These problems are further compounded in cases where the anticipated effect size is small, thus necessitating a larger sample. In such cases, I recommend taking one or more of the three following courses of action. First, when you know that a study lacks statistical power, you should avoid the use of statistical testing. Focus instead on the descriptives, including effect sizes and CIs (see discussion below). Second, in addition to avoiding tests of statistical significance, underpowered studies should also address fewer contrasts between or among groups. For example, if you only expect to be able to recruit 35 participants, rather than comparing four groups/conditions, divide them into two. The additional two conditions can then be compared to themselves and to the first two in a subsequent study. Third, you could bootstrap the analyses or statistics of interest based on the available data/sample (see Larson-Hall & Herrington, 2010; Plonsky et al., in press; LaFlair et al., Chapter 4 in this volume).

However, even if we were able to adequately address the multifaceted “power problem” in L2 research, we would still be relying on the flawed notion of statistical significance. More specifically, a proper understanding and use of statistical power can help the field overcome, at least in part, the unreliability of NHST. The other problems, however, remain. Consider Cumming’s (2012) comments on this issue:

I’m ambivalent about statistical power. On the one hand, if we’re using NHST, power is a vital part of research planning . . . On the other hand, power is defined in terms of NHST, so if we don’t use NHST we can ignore power and instead use precision for research planning . . . However, I feel it’s still necessary to understand power . . . partly to understand NHST and its weaknesses. . . . although I hope that, sometime in the future, power will need only a small historical mention.

(p. 321)

To be clear, I am not suggesting that sample size does not matter. Larger samples will yield less sampling error and, thus, greater precision in our results. The point here, though, is that the notion of statistical power as a means to reliably detect small p values is only relevant within the (flawed) NHST framework. As an

alternative, I argue in the next section that thorough use of descriptive statistics, including effect sizes and CIs, can and should replace much of the statistical testing in L2 research.

Effect Sizes

The focus up to this point in the chapter has been somewhat negative. I have essentially been describing problematic trends and practices in the field. In this section I describe a way forward that helps us to address and improve on these practices by relying on effect sizes in place of NHST. In doing so, I want to address three fundamental questions: (a) What are effect sizes, and how do we calculate them? (b) Why should we use effect sizes? (That is, how is this approach an improvement on current quantitative data practice?) (c) How can we interpret effect sizes?

What Are Effect Sizes, and How Do We Calculate Them?

Let’s start off with a definition of effect sizes: a standardized, quantitative indication of a relationship or an effect. There are many types of effect size indices, but the ones that are most common and applicable in the context of L2 research fall into three categories: mean differences (e.g., d), correlations (r), and variance accounted for (r^2 , R^2 , and η^2).

The first among these, Cohen’s d , is a descriptive statistic that expresses the mean difference between (or within) groups (in SD units—like a z -score). This index is therefore used when we are interested in comparing mean scores, as is often the case in L2 research. The formula for this effect size is very simple:

$$d = \frac{M_1 - M_2}{SD}$$

The difference between means (the numerator) is divided by the pooled standard deviation or that of a control or baseline group, depending on whether the groups have equal variance (see Cumming, 2012). This calculation can be done by hand, but there are also numerous online calculators and Microsoft Excel macros developed for this purpose. (Unfortunately and inexplicably, SPSS does not currently provide Cohen’s d in the output from tests comparing mean scores.) I often use the calculator developed by David B. Wilson that can be downloaded freely here: http://mason.gmu.edu/~dwilsonb/downloads/ES_Calculator.xls. Figure 3.2 shows how user-friendly macros such as this one are. The user simply enters the groups’ means, standard deviations, and sample sizes. The effect size here is $d = .85$, which is based on the sample data I used earlier to show the unreliability of p values. A similar calculator freely available through the Centre for Evaluation and Monitoring is also available here: <http://www.cem.org/effect-size-calculator>. This

calculator has the added advantage of providing CIs around the d value. We can see in Figure 3.3, for example, that the standardized mean difference, which we observed at .85, is likely between .41 and 1.27 in the population. Finally, Hedges’ g , a variant of Cohen’s d , also expresses mean differences and is useful in that it applies a correction for biased effects due to small samples, which are often found in L2 research (Plonsky, 2013).

Though not often viewed this way, correlations such as Pearson’s r are another type of effect size. This index, which ranges from -1 to $+1$, is likely very familiar

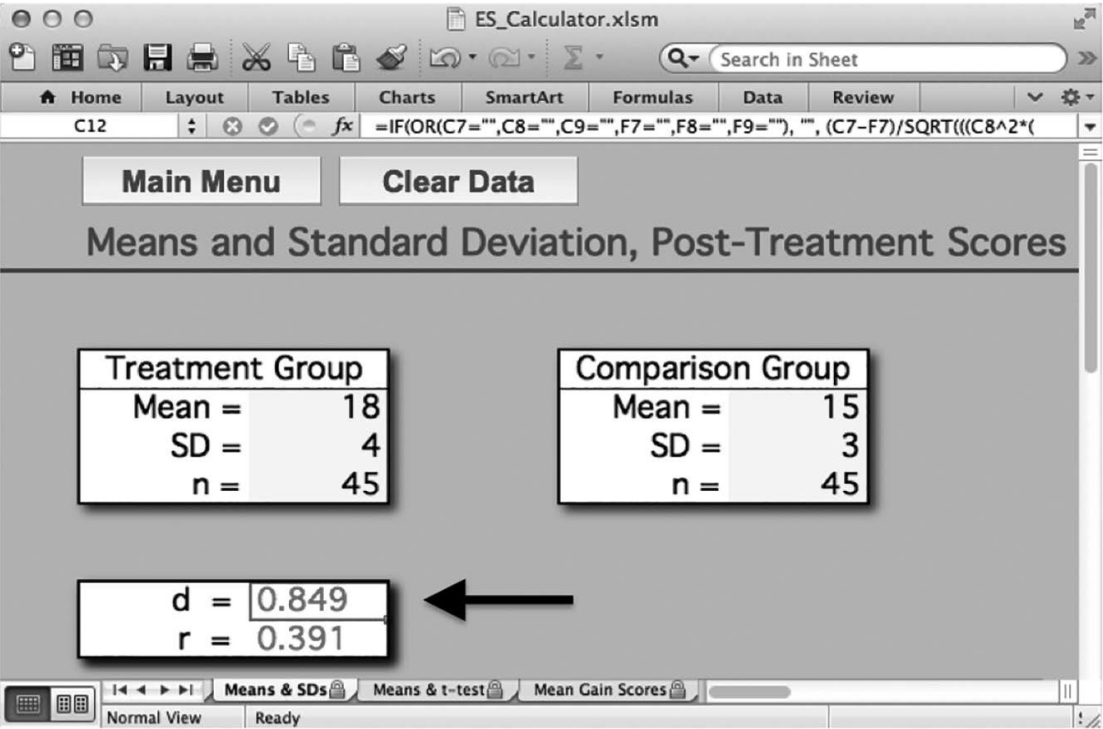


FIGURE 3.2 Screenshot of effect size calculator for Cohen’s d

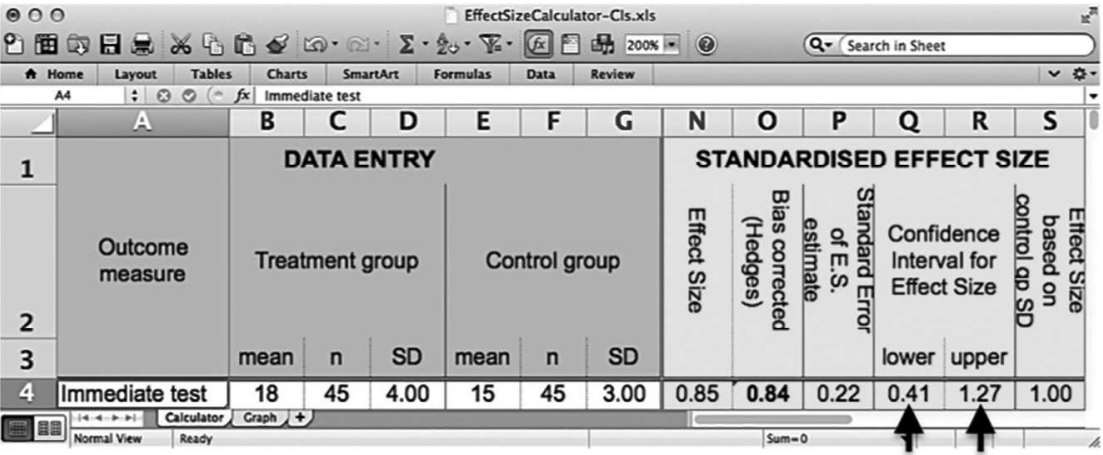


FIGURE 3.3 Screenshot of effect size calculator for Cohen’s d with CIs

to most L2 researchers. There are Web-based applications for calculating correlation coefficients, but most L2 researchers use SPSS. To run a correlation based on normally distributed data, the sequence is **Analyze > Correlate > Bivariate**. You then move your two or more (continuously measured) variables into the Variables box and select *OK*. For example, using the practice data set I’ve made available with this chapter, the correlation between the length (in words) of abstracts and their overall ratings is $r = .38$. (These data are from a study in which Jesse Egbert and I examined the relationship between linguistic and stylistic features of conference abstracts and the scores given to them by raters; Egbert & Plonsky, in press.)

Most researchers reading this are probably very familiar with and used to calculating correlation coefficients. Few, however, are likely aware of how to calculate CIs around this statistic. Again, if we run the correlation described in the previous paragraph, we can see that SPSS does not produce this information automatically, but it can be done by following a short sequence of steps.

The first step is to create new variables based on standardized values of the two variables of interest: **Analyze > Descriptive Statistics > Descriptives**. From within the Descriptives dialogue box, move “Words-tot” and “R_all” into the Variable(s) box. Before clicking *OK*, check the box for *Save standardized values as variables*. The next step is to run the correlation again. However, because we know that SPSS does not produce CIs using the **Correlate > Bivariate** procedure, we have to run the correlation as a simple regression. (You may recall that correlation is simply a type of regression model in which there is a single, continuous predictor variable.) The regression menu can be accessed as follows: **Analyze > Regression > Linear**. Abstract score is our criterion variable so we’ll move our newly created standardized variable for abstract score (“Zscore: R_all”) into the Dependent box on the right. Length is our predictor and we’ll move the standardized variable for length (“Zscore: Words-tot”) to the Independent(s) box. The final command we need to give SPSS is within the Statistics box. Simply click on *Statistics* in the top right corner of the Linear Regression dialogue box, and check the box for *Confidence intervals*. Then click *Continue* to close the Statistics dialogue box and *OK* to run the regression. The two dialogue boxes should look like those in Figures 3.4 and 3.5. The other default settings are fine for our purposes.

The output from this procedure should look like Figure 3.6. We can see in the Standardized Coefficients column that the regression model has produced the same value for the correlation (.38) that we found earlier using the **Correlate > Bivariate** function. This table also provides the 95% CI for that correlation: .272–.488, which tells us the range of values that the true population correlation is likely to fall within. (There are also numerous online calculators that can be used to calculate the CIs for correlation coefficients, such as this one provided by Chris Evans on the PSYCTC website, available at http://www.psyctc.org/stats/R/CI_correln1.html)

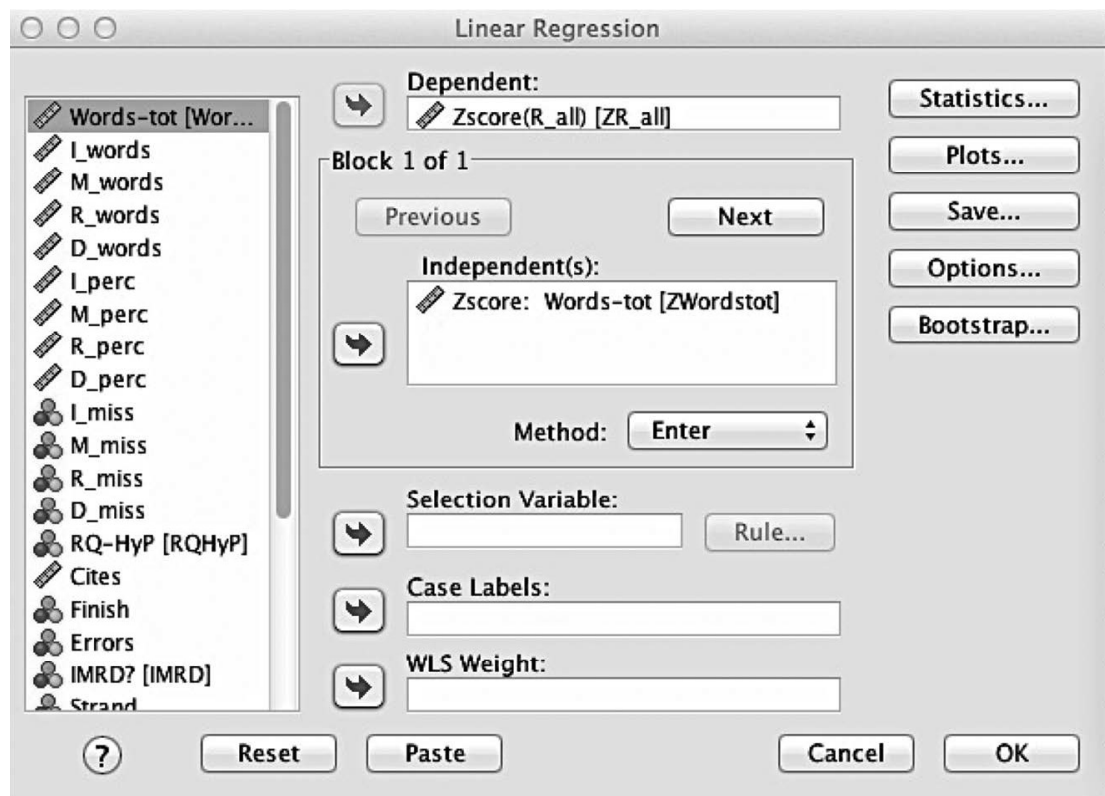


FIGURE 3.4 Linear regression dialogue box used to calculate CIs for correlation coefficients

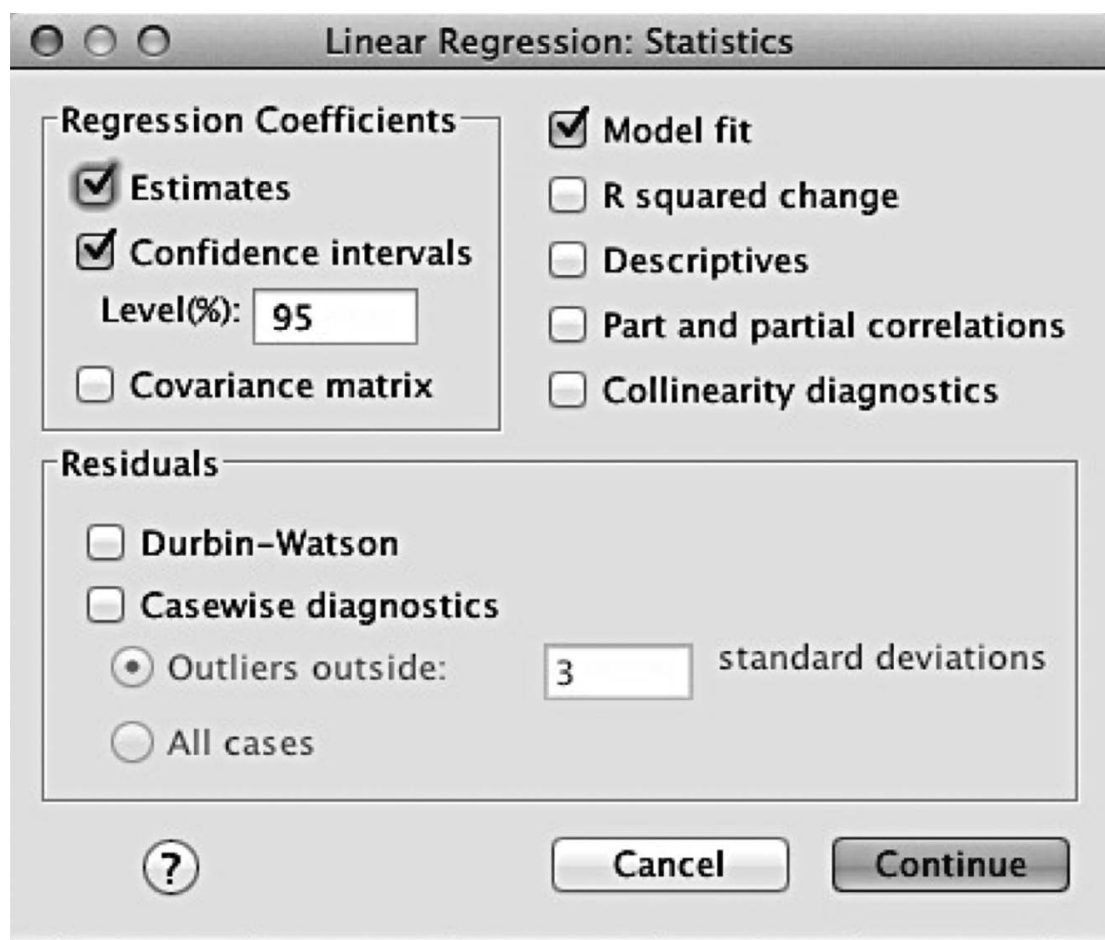


FIGURE 3.5 Statistics dialogue box within linear regression

Coefficients^a

Model	Unstandardized Coefficients		Standardized Coefficients	<i>t</i>	Sig.	95.0% Confidence Interval for B	
	<i>B</i>	Std. Error	<i>Beta</i>			Lower Bound	Upper Bound
1 (Constant)	1.017E-013	0.55		.000	1.000	-.108	.108
Zscore: Words-tot	.380	0.55	.380	6.935	.000	.272	.488

FIGURE 3.6 Output for linear regression with CIs for correlation

Closely related to r , both conceptually and statistically, is a third set or “family” of effect sizes that indicate the extent of shared variance between variables or the amount (%) of variance in one variable that can be accounted for by another. This family includes the R^2 effect size, which we can calculate by simply squaring a correlation coefficient. You’ll recall from the previous example that the correlation we observed between abstract rating and length of words was .38. Once we calculate this value, we can square it ($.38 \times .38$) to determine the amount of shared variance between the two variables: 14%.

In the context of multiple regression analyses (see Jeon, Chapter 7 in this volume), the R^2 effect size index expresses the total or combined variance in the criterion (dependent) variable that is accounted for by the predictor variables. This effect size is produced automatically in the SPSS output for multiple regression. Returning to our abstract study, Egbert and I also used multiple regression to attempt to explain additional variance in abstract ratings. Our model produced a cumulative R^2 value of .31. This result indicates that the set of predictors in our model (e.g., word length, inclusion of results) was able to account for almost a third of the variance in abstract ratings.

The second effect size in this family is eta-squared. You may recognize this effect size as appearing along with ANOVA results and/or in SPSS output. Although we don’t often think of ANOVA as a type of regression, the two procedures are actually quite similar and, consequently, eta-squared, like R^2 , expresses the percent of variance in the dependent variable that can be accounted for by group membership in the independent variable(s). Granena (2013), for example, compared aptitude test scores for native speakers, early L2 learners, and late L2 learners. The results of an ANOVA revealed an eta-squared value of approximately .08. In other words, 8% of the variance in aptitude scores could be accounted for by group membership (i.e., native, early, late). Like r and R^2 , eta-squared can be calculated using SPSS when running ANOVA, but not without asking it to do so. Furthermore, you may have to use a different set of menus than you are used to. Rather than running ANOVA through the Comparing Means menu, to calculate an ANOVA and its corresponding eta-squared value,

you need to run the ANOVA through the General Linear Model drop-down menu: **Analyze > General Linear Model > Univariate**. This procedure will produce an ANOVA. To request an eta-squared value as part of the output, click the *Options* button and check the box for *Estimates of effect size*. An eta-squared value will then be provided in the column labeled as such. Note also that this value for the overall result (“Corrected model”) will be identical to the R^2 value provided as a footnote underneath the output (another remnant of the fact that ANOVA is actually a type of regression, falling under the larger family of general linear models; see Cohen, 1968).

There are several additional types of effect size indices for different types of data and analyses. For categorical or frequency data, researchers may turn to phi and Kramer’s V . Another option for categorical data is a simple percentage. Though not traditionally regarded as an effect size, percentages certainly comply with our earlier definition and, more importantly, they are very easy both to calculate and to interpret. A final effect size commonly used with categorical data is the odds ratio. This index, which expresses the probability of a possible (binary) outcome given a particular condition, is particularly useful in conjunction with logistic regression.

Why Use Effect Sizes?

The main reasons for using effect sizes largely correspond to and address the flaws of NHST described earlier. Recall that the first major flaw was that NHST is unreliable in that any size mean difference or correlation will reach statistical significance given a large-enough sample. Effect sizes, by contrast, are not affected by sample size.² The second major flaw was that NHST is crude and uninformative and that it forces continuous data into a dichotomous (significant/non-significant) result. Furthermore, p values tell us nothing about the extent of the relationship in question (e.g., Cohen, 1994). Effect sizes, however, provide an estimate of the actual strength of the relationship or of the magnitude of the effect in question. Although L2 researchers have been trained, implicitly or explicitly, to set up studies that elicit dichotomous responses, theory and practice can truly be informed only through the more nuanced and precise findings provided in effect sizes. The third and perhaps most obvious flaw of NHST is the arbitrary nature of the .05 cutoff. Unlike p values, effect sizes are continuous, standardized (again, think z -scores), and scale-free. These features of effect sizes enable researchers to make cross-study comparisons and to combine (average) them via meta-analysis.

Beyond these strong conceptual and statistical reasons, I can add one very compelling practical motivation for considering effect sizes: Many major journals now require authors to report them. Following the precedent set by an editorial in *Language Learning* (Ellis, 2000), published in concert with Norris and Ortega’s (2000) seminal meta-analysis in the same issue, several journals that publish L2 research now require authors to report effect sizes. In addition to *Language Learning*, these journals include *Foreign Language Annals*, *Language Learning & Technology*,

Language Testing, *Modern Language Journal*, *Studies in Second Language Acquisition*, and *TESOL Quarterly*. Additionally, many other L2 journals without such explicit policies adhere to APA style, which also requires reporting of effect sizes.

As a result of both the benefits described and the relatively recent requirements of journals in this area, the presence of effect sizes has increased substantially in recent years. Plonsky and Gass's (2011) review of methodological practices in the interactionist tradition found, for example, that whereas none of the 174 studies they examined in the 1980s or 1990s reported effect sizes, 27% of the studies published in the 2000s did so. Likewise, Plonsky (2014) found the percentage of studies reporting effect sizes to increase exponentially from 3% in the 1990s to 42% in the 2000s.

Interpreting Effect Sizes

It is clear that the field's interest in effect sizes is increasing. However, primary researchers currently do little in the way of using effect sizes to enhance our results or, more importantly, our understanding of the variables and relationships we study. (The same could be said for descriptive statistics more generally.) That is, most authors currently treat effect sizes as a hoop to jump through or box to check as part of a set of manuscript submission guidelines. What authors need to do is provide more meaningful interpretations of the full range of descriptives in their data, including of course their effect sizes.

Unlike p values, which are usually understood in a very straightforward (but equally uninformative) manner (i.e., significant/nonsignificant), effect sizes require more nuanced interpretation. Taking advantage of the rich information provided by effect sizes forces us to address questions such as “What does a d value of .65 mean (for theory and practice)?” “What constitutes a small or large effect in this particular domain?” and “How does a correlation of, say, .35 compare with the predictions of theory for the relationship between these two variables?”

There are a number of different approaches for addressing these questions (see Stukas & Cumming, in press). One very common approach has been to compare observed effects to benchmarks designed for this purpose. Based on their synthesis of effects from 346 primary studies and 91 meta-analyses ($N > 604,000$), Plonsky and Oswald (2014) proposed a general scale for interpreting d and r values in L2 research (Table 3.5). Values for each type of effect size, labeled as roughly small, medium, and large correspond approximately to the 25th, 50th, and 75th percentiles of observed effects in their sample. Such benchmarks can be useful as a means to situate the effects of a particular study in relation to the larger field. The authors also caution, however, that doing so should only be considered a first step in the interpretation of effect sizes. In other words, we cannot assume that what constitutes a large effect in one area of L2 research is necessarily the same as what one would expect to be a large effect in all other areas.

TABLE 3.5 General benchmarks for interpreting d and r effect sizes in L2 research

<i>Effect size</i>	<i>Small</i>	<i>Medium</i>	<i>Large</i>
Mean difference (d)			
Between-groups	0.40	0.70	1.00
Within-groups	0.60	1.00	1.40
Correlation (r)	0.25	0.40	0.60

Indeed, there are a number of additional factors that merit consideration when interpreting effect sizes. Most critically, researchers must provide an explanation of what the particular numerical effects they observe mean in the context of their domain. Others factors, discussed at length in Plonsky and Oswald (2014), include (a) effects found in previous studies in the same subdomain; (b) mathematical readings of effect sizes (see Plonsky & Oswald, 2014, pp. 893–894); (c) theoretical and methodological maturity of the domain in question; (d) research setting (e.g., lab vs. classroom); (e) practical significance; (f) publication bias in previous research; (g) psychometric properties and artifacts; and (h) other methodological features.

Descriptive Statistics: Means, Standard Deviations, and CIs

In addition to calculating and interpreting effect sizes, it is absolutely critical that researchers become very familiar with their descriptive statistics. (I realize this will sound obvious to many of you, but scholars in our field often carry out statistical tests without ever first conducting a thorough examination of their descriptive statistics.) In the case of research investigating mean differences, those means are probably a good place to start. But they are just that: a starting point. Mean scores give an initial indication of the difference(s) between two or more groups. They say nothing, however, about the spread of scores around them. For this crucial information, we usually look at standard deviations.

I have to point out here that the importance of understanding the spread of scores can hardly be overstated. This concept, called *variance*, is deeply entrenched in nearly all statistical techniques employed in L2 research and across the social sciences (see GLM in Plonsky, Chapter 1, p. 5). For example, though we tend to think of ANOVA (analysis of variance) as a comparison of means, it is just as much if not more concerned with within- and between-group variance. Recall from the previous section that a standard deviation forms the denominator in the formula for Cohen's d . Despite the centrality of this statistic and the concept it represents, very rarely do L2 researchers give any explicit consideration of standard deviations in written reports. In fact, it is quite common for them to be left out of published L2 research (e.g., Plonsky, 2013).

In terms of statistical testing and comparisons of mean scores, when there is a lot of variance (large standard deviations) group scores are more likely to overlap. Consequently, the results of a *t* test or ANOVA are less likely to be statistically significant and their corresponding *d* values will be smaller. More conceptually speaking, a close look at the standard deviation can help you decide how much faith to put in your mean with respect to its ability to represent your sample. Standard deviations can also provide insights into substantive issues. For example, in experimental designs, an increase in the experimental group’s standard deviation from pre- to posttest might indicate that not all learners responded uniformly to the treatment and that there may be learner-internal moderators at play.

A related descriptive statistic that is considered and reported even less frequently is the CI. CIs express a range of values around an observed mean score that are likely (at a given level of probability, typically 95%) to contain the true population mean. Returning to the abstract study described earlier, imagine you were interested in understanding typical abstract ratings. We might begin by calculating the mean score for this variable along with its corresponding 95% CIs and other descriptives. The series of commands using SPSS is as follows: **Analyze > Descriptive Statistics > Explore**. (See steps for calculating CIs for correlations above.) From there we simply move the “R_all” variable into the Dependent list. The CIs are set at 95% by default, but if you had reason to set them more strictly or more leniently, you could do so using the Statistics dialogue box. After clicking OK, the resulting output shown in Figure 3.7 would provide a full set of descriptive statistics including the CIs. (This is one reason I almost never calculate descriptives using SPSS using the **Analyze > Descriptive Statistics > Descriptives** menu—it is not nearly as informative as the **Explore** function.)

Calculating CIs and other descriptives using Excel is also quite simple:

1. Calculate the mean score by typing in the following in the first empty cell at the bottom of the column of data you are interested in: `=AVERAGE(X:Y)`, where X and Y refer to the top and bottom cells of data (be sure to exclude any header rows).
2. In the cell immediately below the mean score, calculate the standard deviation for the set of scores: `=STDEV(X:Y)`, where X and Y are the same as for the step 1.
3. In the cell immediately below the standard deviation, calculate the interval that will be added and subtracted from the mean score to construct the CI: `=CONFIDENCE.NORM(alpha,SD,N)`. The *alpha* field here is usually .05, corresponding to a 95% CI, but could easily be adjusted; for a 90% CI, for example, this value would be .1. In the *SD* field of this formula, simply type in the name of the cell where that value was calculated in step 2 (e.g., U55). And the *N* field refers to the number of data points/cases/observations in the sample.

4. Construct the upper and low bounds of the CI by adding/subtracting the value from step 3 to/from the mean calculated in step 1. Simply type into two empty cells: $= B - C$ and $= B + C$, where B refers to the mean score calculated in step 1 and C refers to the interval calculated in step 3, respectively.

There are many ways to interpret CIs (see Cumming, 2012), but their primary purpose is to help us situate mean scores in the context of the many other possible values that might represent the true population score (as opposed to that of the sample). As Carl Sagan (1996) put it, CIs are “a quiet but insistent reminder that no knowledge is complete or perfect” (pp. 27–28). As with standard deviations, considering the CIs around our mean scores, numerically and/or visually, helps us avoid the temptation to view our samples and their mean scores as absolute.

In the case of abstract ratings for this particular L2 research conference, we can see in Figure 3.7 that the mean score is 3.64 (on a scale of 1–5) with 95% CIs of [3.56, 3.71]. (CIs are typically reported in brackets.) The width of the interval is quite narrow, which is likely due to the relatively large sample ($N = 287$). Consequently, assuming these data are based on a valid and reliable instrument, we can be fairly confident that our point estimate of 3.64 is very close to the true population mean for scores at this conference.

CIs can also be used to indicate whether the difference between a pair of mean scores is statistically significant and whether that difference is stable. This information is also quite easy to access: We simply check to see whether the mean of one group falls within or outside the CI for the other group’s mean. We can try this out using the abstract data set. Let overall score here be the dependent variable and let the presence of one or more errors be a dichotomous independent

Descriptives

			Statistic	Std. Error
R_all	Mean		3.6359	.03923
	95% Confidence Interval for Mean	Lower Bound	3.5587	
		Upper Bound	3.7131	
	5% Trimmed Mean		3.6568	
	Median		3.7500	
	Variance		.442	
	Std. Deviation		.66464	
	Minimum		1.75	
	Maximum		5.00	
	Range		3.25	
	Interquartile Range		1.00	
	Skewness		-.441	.144
	Kurtosis		-.316	.287

FIGURE 3.7 Output for descriptive statistics produced through Explore in SPSS

variable. The menu sequence using SPSS is, again, **Analyze > Descriptive Statistics > Explore**. This time, however, we will move the “Errors” variable into the “Factor list” box. As we can see in Figure 3.8, the mean score for the “no errors” group (3.68) does not fall within the CI for the “error(s) present” group [3.23, 3.60] and vice versa, thus indicating that the difference between these two means is statistically different. We can also calculate the effect size for the difference between these groups using one of the tools described earlier: $d = .40$.

Though it is not strictly necessary, we could confirm this result by running an independent samples t test, which would produce a t value of 2.62 with an associated p value of .009. An advantage to following up our analysis based on CIs with a t test is that the SPSS output will also provide a CI around the mean difference, which can help us better understand how stable it is. In this particular case, the mean difference between the two groups is .26, and the CI associated

Descriptives

<i>Errors</i>		<i>Statistic</i>	<i>Std. Error</i>
R_all	no errors	Mean	3.6837
		95% Confidence Lower Bound	3.5990
		Interval for Mean Upper Bound	3.7684
		5% Trimmed Mean	3.7071
		Median	3.7500
		Variance	.435
		Std. Deviation	.65933
		Minimum	1.75
		Maximum	5.00
		Range	3.25
		Interquartile Range	1.00
		Skewness	-.506
		Kurtosis	-.188
			.159
			.316
	error(s) present	Mean	3.4199
		95% Confidence Lower Bound	3.2385
		Interval for Mean Upper Bound	3.6013
		5% Trimmed Mean	3.4231
		Median	3.5000
		Variance	.425
		Std. Deviation	.65158
		Minimum	2.17
		Maximum	4.75
		Range	2.58
		Interquartile Range	.90
		Skewness	-.236
		Kurtosis	-.537
			.330
			.650

FIGURE 3.8 Descriptive statistics and CIs for abstracts with vs. without errors

with that difference is [.07, .46]. Yet another confirmation of the statistical difference between these means scores here is the fact that the CI around the mean difference does not cross 0. What is perhaps more interesting is to note that the CI is somewhat narrow, indicating that our point estimate for the difference (.26) is rather stable and reliable. If the CI had been much larger relative to the five-point scale, say [.20, 3.9], we would have less certainty—that is, confidence—in our observed mean difference. For a number of worked examples and practice interpreting CIs, see Cumming (2012) and, in the context of L2 research, Larson-Hall and Plonsky (2015, p. 135).

Finally, it is not sufficient to simply calculate and examine a full set of descriptive statistics when analyzing quantitative data. Such results also need to be made available in published reports and/or appendices to justify interpretations and to enable consumers of L2 research to draw their own conclusions as well. More complete reporting of data also assists in meta-analyses and other synthetic efforts. For these reasons and in line with the APA (2010), all mean-based analyses should be reported, at a minimum, with their associated means, standard deviations, CIs, and effect sizes (again, see Larson-Hall & Plonsky, 2015).

Looking Forward

The impetus behind this chapter—the entire volume, really—is to improve and advance L2 research practices. Toward that end, I'd like to propose a *revised* model of L2 research (Figure 3.9) both as a point of contrast with the descriptive model in Figure 3.1 and as a suggestion for how our individual and collective research efforts ought to proceed.

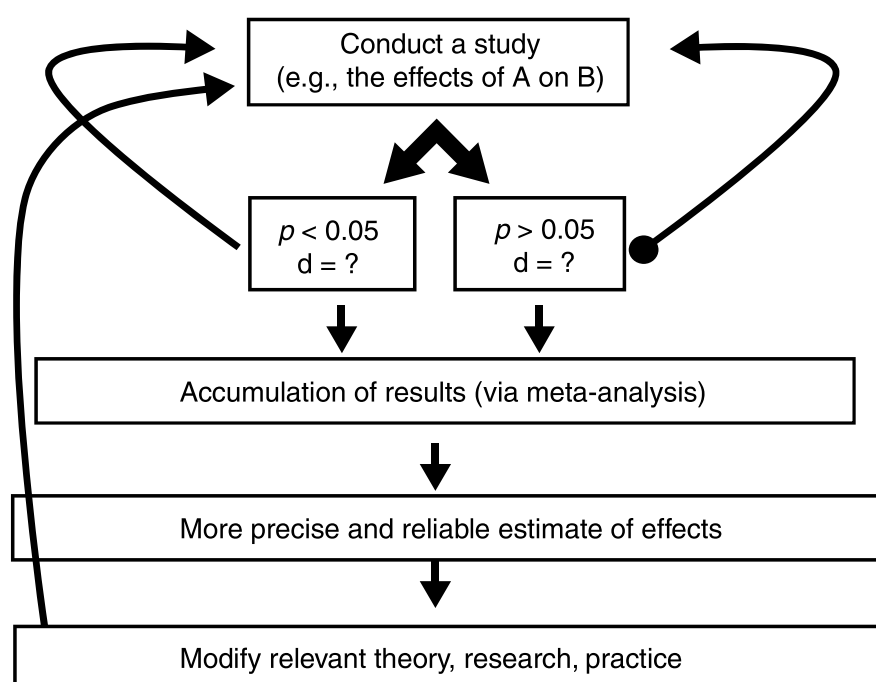


FIGURE 3.9 A revised model of quantitative L2 research

As with the model currently in place, the process begins when a researcher conducts a study. Unlike the current model, however, assuming the study is well designed, the importance of the study’s findings and its likelihood of getting published do not hinge on the flawed notion of statistical significance. Rather, both statistical and practical significance are considered and interpreted, and the results of the study and others in the domain are brought together via research synthesis and meta-analysis. By embracing a synthetic research ethic both at the primary and secondary levels, the domain in question is able to arrive at a view of the relationships or effects in question that is more reliable, thereby enabling L2 theory and practice to be more accurately informed by empirical efforts.

Tools and Resources

The following links provide access to very user-friendly programs for conducting many of the analyses described in this chapter. The first, the langtest.jp developed by Atsushi Mizumoto, is an R- and web-based app (<http://langtest.jp/>); the second, ESCI (<http://www.latrobe.edu.au/psy/research/cognitive-and-developmental-psychology/esci>), is a set of freely downloadable Excel macros designed by Geoffrey Cumming to help researchers consider and report results with an emphasis on effect sizes and CIs.

Further Reading

- Cohen, J., (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.
- Cumming, G. (2012). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. New York: Routledge.
- Kline, R.B. (2013). *Beyond significance testing: Statistics reform in the behavioral sciences* (2nd ed.). Washington, DC: American Psychological Association.
- Larson-Hall, J. (2010). *A guide to doing statistics in second language research using SPSS*. Chapter 4. New York: Routledge.
- Wilkinson, L., & Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, 54, 594–604.

Discussion Questions

1. Summarize, in your own words, the main arguments against the use of p values and, conversely, in favor of “estimation thinking” and effect sizes. Can you think of any counterarguments or situations in which the NHST approach might be preferable or even justifiable?
2. Considering the current place of NHST and effect sizes in quantitative L2 research, what changes would you suggest to the field?
3. The subtitle of this chapter (“A back-to-basics approach to advancing quantitative methods in L2 research”) implies that power and statistical vs. practice significance have been around for a while. If this is the case, why have we as a field been so slow to embrace these notions in these research practice?

4. Find a quantitative study in your area of interest. To what extent does it adhere to NHST and associated data analytic techniques and interpretations? How could the study be revised to provide more informative and precise results?

Notes

1. These values also assume a normal distribution; variance must also be considered in calculating power and effect sizes.
2. However, the width of CIs for effect sizes is influenced by sample size.

References

- Cohen, J. (1968). Multiple regression as a general data-analytic system. *Psychological Bulletin*, 70, 426–443.
- Cohen, J. (1992). A power primer. *Psychological Bulletin*, 112, 155–159.
- Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist*, 49, 97–1003.
- Cumming, G. (2012). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. New York: Routledge.
- Cunnings, I. (2012). An overview of mixed-effects statistical models for second language researchers. *Second Language Research*, 28, 369–382.
- Egbert, J., & Plonsky, L. (in press). Success in the abstract: Exploring linguistic and stylistic predictors of conference abstract ratings. *Corpora*.
- Ellis, N.C. (2000). Editorial statement. *Language Learning*, 50, xi–xiii.
- Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, 39, 175–191.
- Granena, G. (2013). Individual differences in sequence learning ability and second language acquisition in early childhood and adulthood. *Language Learning*, 63, 665–705.
- Kline, R.B. (2013). *Beyond significance testing: Statistics reform in the behavioral sciences* (2nd ed.). Washington, DC: American Psychological Association.
- Larson-Hall, J., & Herrington, R. (2010). Improving data analysis in second language acquisition by utilizing modern developments in applied statistics. *Applied Linguistics*, 31, 368–390.
- 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, 125–157.
- Loewen, S., Lavolette, B., 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.
- Mackey, A., & Sachs, R. (2012). Older learners in SLA research: A first look at working memory, feedback, and L2 development. *Language Learning*, 62, 704–740.
- Norris, J. M. (in press). Statistical significance testing in second language research: Basic problems and suggestions for reform. *Language Learning*, 65, Supp. 1.
- Norris, J. M., & Ortega, L. (2000). Effectiveness of L2 instruction: A research synthesis and quantitative meta-analysis. *Language Learning*, 50(3), 417–528.

- 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). Amsterdam: John Benjamins.
- Oswald, F.L., & Plonsky, L. (2010). Meta-analysis in second language research: Choices and challenges. *Annual Review of Applied Linguistics*, 30, 85–110.
- 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., & Gass, S. (2011). Quantitative research methods, study quality, and outcomes: The case of interaction research. *Language Learning*, 61, 325–366.
- Plonsky, L., Egbert, J., & LaFlair, G.T. (in press). Bootstrapping in applied linguistics: Assessing its potential using shared data. *Applied Linguistics*.
- Plonsky, L., & Oswald, F.L. (2014). How big is ‘big’? Interpreting effect sizes in L2 research. *Language Learning*, 64, 878–912.
- Porte, G. (2010). *Appraising research in second language learning: A practical approach to critical analysis of quantitative research* (2nd ed.). Philadelphia/Amsterdam: John Benjamins.
- Rosnow, R.L. & Rosenthal, R. (1989). Statistical procedures and the justification of knowledge in psychological science. *American Psychologist*, 44, 1276–1284.
- Sagan, C. (1996). *The demon-haunted world*. New York: Random House.
- Stukas, A.A., & Cumming, G. (in press). Interpreting effect sizes: Towards a quantitative cumulative social psychology. *European Journal of Social Psychology*.
- Taylor, A.M., Stevens, J.R., & Asher, J.W. (2006). The effects of explicit reading strategy training on L2 reading comprehension: A meta-analysis. In J.M. Norris & L. Ortega (Eds.), *Synthesizing research on language learning and teaching* (pp. 213–244). Amsterdam: John Benjamins.
- Thompson, B. (1992). Two and one-half decades of leadership in measurement and evaluation. *Journal of Counseling and Development*, 70, 434–438.