

CHAPTER 6

Reporting and Interpreting Quantitative Research Findings: What Gets Reported and Recommendations for the Field

Jenifer Larson-Hall^a and Luke Plonsky^b

^aFukuoka Jo Gakuin and ^bNorthern Arizona University

This paper presents a set of guidelines for reporting on five types of quantitative data issues: (1) Descriptive statistics, (2) Effect sizes and confidence intervals, (3) Instrument reliability, (4) Visual displays of data, and (5) Raw data. Our recommendations are derived mainly from various professional sources related to L2 research but motivated by results from investigations into how well the field as a whole is following these guidelines for best methodological practices, and illustrated by L2 examples. Although recent surveys of L2 reporting practices have found that more researchers are including important data such as effect sizes, confidence intervals, reliability coefficients, research questions, a priori alpha levels, graphics, and so forth in their research reports, we call for further improvement so that research findings may build upon each other and lend themselves to meta-analyses and a mindset that sees each research project in the context of a coherent whole.

Keywords confidence intervals; effect sizes; language; L2; mean; meta-analysis; methodology; quantitative; raw data; reporting practices; standard deviation

Introduction

Thorough and consistent reporting practices are critical to advancing and informing second language (L2) theory and practice. In this chapter we call for quantitative L2 researchers to adopt a synthetic mindset—understanding that no research stands alone, that progress as a field happens best if we can pool research findings, and that reporting practices can either facilitate or impede

Correspondence concerning this article should be addressed to Jenifer Larson-Hall, Fukuoka Jo Gakuin, Centre for the Study of English Language Teaching, Fukuoka-shi, Minami-ku, Osa 3-42-1. E-mail: drlarsonhall@gmail.com

this type of research. However, currently researchers have little in the way of L2-specific guidance or direction on what to present and how. As a baseline from which to assess L2 quantitative research, we will first introduce reporting practices based on recommendations from various sources related to L2 research. Next, we will review how well the field as a whole follows these guidelines by presenting results from secondary/meta-analytic research. We then propose a set of recommendations for improving the means by which quantitative L2 data are reported and interpreted. We do this for five main areas of quantitative reporting: (1) Descriptive statistics and other measures of study quality such as power, (2) Effect sizes and confidence intervals (CIs), (3) Instrument reliability, (4) Visual displays of data, and (5) Raw data.

Introduction to Reporting Standards

A number of scholars have raised concerns over the years with respect to reporting practices in quantitative L2 research (e.g., Chaudron, 2001; Henning, 1986; Norris & Ortega, 2000; Nunan, 1991). The reality of L2 research is that it is extremely heterogeneous and draws upon or is influenced by practices in a number of different domains, including Psychology, Education, and Linguistics. In order to help illuminate a path toward more standardized and transparent reporting procedures in the L2 field, we begin by reviewing existing frameworks for data reporting standards, which we then use as the basis for our recommendations at the end of the chapter.

Many subdomains of L2 quantitative research use various types of empirical methodologies, including controlled experiments, quasi-experiments (where assignment to groups is not random), surveys and questionnaires, observational studies, case studies, statistical simulations, and meta-analyses. Although there are various study approaches to research questions, most of these types of empirical research report numbers and will thus be considered quantitative types of research reports.

Many L2 researchers rely at least in part on the reporting standards of the publication manual of the American Psychological Association (APA, 2010). This comprehensive guide provides explicit direction on the types of data that should be reported in quantitative research. For example, the APA manual prescribes reporting of “all relevant results, including those that run counter to expectation” (p. 32), effect sizes (again, for all findings, not only those that confirm researcher expectations or that reach statistical significance), CIs, standard deviations (*SDs*) to accompany means, exact *p* values, exact test statistics, and degrees of freedom. The APA manual also recommends that, with

respect to the interpretation of data, the majority of the discussion be based on point and interval estimates (i.e., effect sizes and CIs) rather than the binary outcomes of statistical tests (i.e., significant or not). In addition, the manual compels authors to share their data with other researchers for the sake of re-analysis, which can be very useful but which may come as a surprise to many L2 researchers. (See Plonsky, Ebgert, & LaFlair, in press, for a report on applied linguists' openness to sharing data.) The APA has also produced several other useful publications that specifically address and provide guidance on data reporting. These include the APA's (2008) *Journal Article Reporting Standards* and three additional book-length treatments of reporting practices covering (a) table design (Nicol & Pexman, 2010b); (b) figures, posters, presentations (Nicol & Pexman, 2010a); and (c) a general guide to data reporting (Cooper, 2011).

Other types of common L2 research may draw more on guidelines from other fields, such as education. In 2006, the American Educational Research Association (AERA), for example, published their own set of reporting standards for empirical research. Their guide provides numerous recommendations for study design, implementation, and data reporting, many of which align very closely with those of the APA. For example, the report emphasizes the importance of reporting both the size of and uncertainty in quantitative measures, meaning that descriptive statistics such as means and standard deviations should always be provided. For inferential statistics, the report outlines the importance of effect sizes and recommends CIs along with standard deviations as a way to indicate the uncertainty in effect sizes or point estimates. It also echoes the APA in calling for researchers to interpret quantitative findings in terms of practical usefulness for the research question at hand.

Taking notice of the above recommendations, which provide a confluence of ideas, we propose setting L2 standards for reporting in the five areas of descriptive statistics, effect sizes and CIs, reliability reports, visual displays of data, and raw data. We will consider each of these areas in turn (drawing upon the APA recommendations as the primary basis for our recommendations), look at reports of how well the field currently adheres to them, and summarize the results at the end of the paper.

At this point, however, the reader will likely be posing the very common question of whether researchers will have enough room to report the information we are calling for. The history of L2 research shows somewhat of an evolution toward greater sophistication in statistical use and reporting (Loewen & Gass, 2009), but traditional limits on page length may have been impediments to full reporting of statistical details. Research papers, often commonly thought of as

approximately 30 pages long, have more recently become longer as more details about many different parts of the research report are asked for, with Catford (1998) reporting nearly two decades ago that the average article in *Language Learning* was running 40 pages. However, today there are more alternatives to work around page limits, such as online appendices to supplement print journals as well as online-only journals where space considerations are not as important. We would also suggest that much of what we are arguing for here can be reported using very little additional space. Including and interpreting CIs in a report, for example, will usually entail adding an extra column to one or more tables and perhaps just a few additional lines of text at most.

Descriptive Statistics and Measures of Study Quality

Descriptive statistics are the basis of all quantitative reasoning and it is absolutely necessary that they be reported in research studies. They are fundamentally essential to everything else we do quantitatively because they describe the data set, providing a point estimate of the average trend of the data (the mean, or more rarely in our field, the median) as well as estimates of variability in the data (e.g., the standard deviation). The mean and standard deviation (*SD*) provide the foundation upon which most statistical calculations are based and should not be ignored in favor of those statistical conclusions—indeed, they should be used in explaining statistical findings. Other study features and data that ought to be present in published reports for various reasons (all of which are required or encouraged by APA) include (a) research questions and/or hypotheses, (b) an a priori alpha level (when inferential statistics are used), (c) the statistical assumptions that were checked (and, if unmet, any accommodations made), (d) whether statistical power or precision (if CIs are being used) was considered in the study's design in order to determine an appropriate sample size, (e) exact *p* values (when inferential statistics are used), and (f) additional descriptive statistics (e.g., minimum and maximum scores, numbers of items on instruments, numbers of participants or observations in all groups or samples).

It is vital that basic descriptive statistics such as mean (or where appropriate, median) scores as well as measures of variability such as *SDs* be reported in research reports. Many L2 researchers may realize that they need to report means and *SDs*, but fail to report them for all the configurations of their data that they analyze. For example, imagine an analysis of variance (ANOVA) design where the researcher wanted to explore the effects of L1 on L2 pronunciation and hypothesized that word class (noun, verb, adjective) affected pronunciation. The researcher might report means and standard deviations for each L1 group

as a whole and find in the ANOVA that there was a difference between L1 groups, but then go on to conduct pairwise comparisons between various L1 groups on word categories without reporting means and standard deviations for performance for each group on each word category. It is essential that means (or medians) and standard deviations be reported not just once for each data set but also for each grouping of the data that is of interest in the research.

When descriptive statistics are omitted in research reports, this is not only bad reporting practice that goes against the APA guidelines, but a practice that harms our field as a whole. If an individual researcher does not report descriptive statistics for their data (the primary level), then that information cannot be used later in research synthesis and meta-analysis (the secondary level of data analysis). A meta-analysis is a set of techniques that seeks to collect all of the primary data about a particular issue (such as the effect of explicit vs. implicit instruction, or the effectiveness of visual presentation of vocabulary) and combine the quantitative results to provide a more comprehensive picture of the effects of a treatment or condition. Cumming (2012) suggests that “at its simplest, it [the meta-analysis] gives a point estimate that is a weighted average of the separate study means. It also gives an overall interval estimate that signals how precise an estimate the weighted average is likely to be” (p. 5). For example, in an early L2 meta-analysis conducted by Norris and Ortega (2000), which synthesized primary research to try to answer a number of questions about the effectiveness of L2 instruction, they found that studies that investigated the effectiveness of a Focus on Form¹ treatment had a mean point estimate of Cohen’s $d = 1.00$ (meaning the difference between groups was on the magnitude of one standard deviation), with a 95% CI estimate for Cohen’s d of [0.78, 1.22] (meaning that we are 95% confident that this interval includes the actual mean value). In contrast, studies that looked at the effect of Focus on Form using implicit methods only had a mean level of effectiveness of $d = .069$, 95% CI [0.38, 1.00], while studies using explicit methods had a higher mean of $d = 1.22$, 95% CI [0.91, 1.53]. This particular question examined data from 43 primary studies and the results helped the field as a whole understand how effective a Focus on Form is; however, it was only possible to come to these findings based on studies that reported the necessary descriptive statistics for calculating effect sizes and CIs. Clearly, when primary research does not report descriptive statistics, such studies must often be excluded from meta-analyses, thus limiting their contribution to the advancement of knowledge.

Meta-analytic thinking thus involves understanding that the results from any one study will not be conclusive and that to draw solid conclusions, data from a number of studies need to be gathered and analyzed. Cumming (2012)

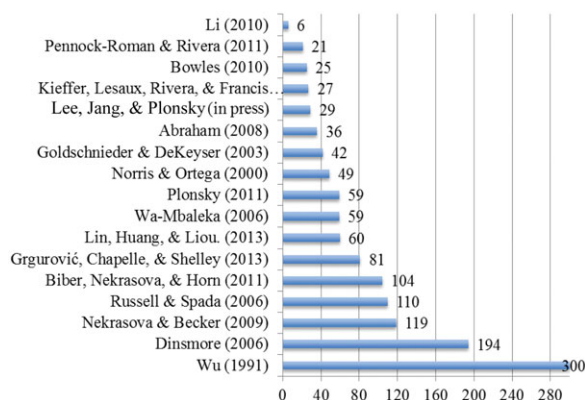


Figure 1 Reports from L2 meta-analyses that counted the number of studies excluded from the analysis due to missing data (expressed as a % of the included sample).

states that “[m]eta-analytic thinking is the consideration of any result in relation to previous results on the same or similar questions, and awareness that combination with future results is likely to be valuable. . . . It prompts us to seek meta-analysis of previous related studies at the planning stage of research, then to report our results in a way that makes it easy to include them in future meta-analyses” (p. 9).

Because meta-analysts rely almost exclusively on the reporting practices of primary researchers, missing and/or unreported data at the primary level necessarily yields gaps in our knowledge at the secondary or meta-analytic level. This problem appears somewhat widespread in L2 research as evidenced by data from numerous L2 meta-analyses. In this and all subsequent examples, meta-analyses from the approximately 90 L2 analyses that were known to the second author at the time of reporting that provided the relevant information were consulted in order to provide examples and evidence. Figure 1 shows the number of studies excluded from 17 meta-analyses of L2 research due to missing data. The number of studies that were excluded from each meta-analysis is expressed as a percentage of the number of studies included in the final sample, which means that the number can be larger than 100%. For example, if there were 200 research reports that examined data from interactionist studies but only 50 could be included in the final summary due to missing data in 150 other reports, the exclusion rate would be 300%. In the vast majority of cases, these studies were excluded because insufficient data were reported to calculate an effect size—usually *SDs*. The range of percentages of studies excluded for this reason ranges widely from 6–300% with a median of 59%. To be clear,

these data indicate that omitted statistics—or, more precisely, the authors who omitted them—are responsible in some cases for rendering massive amounts of data un-meta-analyzable and therefore unavailable to contribute to already limited efforts to aggregate findings across studies. Such field-wide gaps are extremely inefficient when accumulation of knowledge is the goal.

Some meta-analyses of L2 research have also indicated the extent to which specific kinds of descriptive and inferential statistics were omitted in their samples. Exact means for the specific groups being compared, for example, were left out of 14% (Nekrasova & Becker, 2009), 18% (Norris & Ortega, 2000), 20% (Plonsky, 2013), and 29% (Keck, Iberri-Shea, Tracy-Ventura, & Wa-Mbaleka, 2006) of the samples meta-analyzed by L2 researchers. Likewise, *SDs* were omitted in 6% (Wang, 2010), 12% (Plonsky & Gass, 2011), 29% (Keck et al., 2006), and over 50% (Norris & Ortega, 2000) of other meta-analyzed samples. Plonsky's (2013) findings across L2 subdomains paint an even more disturbing picture. His synthesis found that 31% of the sample (or higher, depending on how you divide up the domains) failed to report one or more *SDs* along with their corresponding means. This pattern is indicative of a lack of understanding that means are almost meaningless without some measure of their variability and is also not good for meta-analyses, which require *SDs* to calculate effect sizes. Those familiar with effect sizes may rightly point out that Cohen's *d*, the effect size index in nearly all these meta-analyses, can be calculated based on test statistics such as the *t* value resulting from a *t* test. Unfortunately, even exact *t* and *F* values are often left out of primary reports as well: Specifically, this was the case in 17% (Norris & Ortega, 2000), 19% (Lee & Huang, 2008), 24% (Plonsky, 2013), and 49% (Wang, 2010) of these meta-analyzed samples. Exact *p* values, which can also be used to calculate *d* values in the absence of descriptive statistics, were present in only 44% (Plonsky & Gass, 2011) and 49% (Plonsky, 2013) of the sampled research. But the latter study found that 42% of the sample also did so inconsistently. In other words, *p* values were often found in a single paper to be reported as both an exact value and relative to a cutoff such as .05.

In looking at other measures of basic reporting standards, Plonsky (2013) found frequent reporting of research questions—80% of the 606 articles surveyed. However, other basic features were found more sporadically. Stating a predetermined level of statistical significance such as .05 is not common, and several meta-analyses found that pre-set alpha levels were only reported about 25% of the time (Keck et al., 2006; Mackey & Goo, 2007; Norris & Ortega, 2000; Plonsky, 2013; Plonsky & Gass, 2011; Ziegler, 2013). Similarly, Crookes's (1991) assertion that power analyses “are almost never used” is still

true today (p. 762). Studies that examined whether power analyses were reported found that only a very small percentage of studies did so (2% in Plonsky & Gass, 2011; 1% in Plonsky, 2013, and 7% in Ziegler, 2013). The proportion of studies that has reported checking statistical assumptions prior to employing inferential statistics is similarly small (3% in Plonsky & Gass, 2011; 17% in Plonsky, 2013, and 36% in Ziegler, 2013), a result we find disturbing given the high co-occurrence of parametric statistics on the one hand and non-normal distributions, small samples, and outliers on the other (Larson-Hall, 2010; Larson-Hall & Herrington, 2009; Phakiti, 2010; Plonsky, 2013; Larson-Hall, 2015a; Plonsky et al., in press).

Beyond simply reporting descriptive statistic such as means, *SDs*, and exact test statistics and *p* values as well as a priori power analyses, we urge researchers to interpret these numbers as well. For example, variability (often expressed as a *SD*) is an area of great informational richness—indeed, it is the basis for the inferential techniques, like ANOVA so favored by L2 researchers—but it is almost never discussed. Readers may wonder what they would say about their *SDs*, and here is where a meta-analytic/synthetic mindset is the goal we are urging readers toward. *SDs* could be interpreted in light of the groups involved in the study being reported on, but even more illuminating might be comparisons with *SDs* in other, similar studies. To summarize this section, we would like to urge the field toward reporting and interpreting results in the context of a larger line of inquiry and the understanding that no single study can provide a conclusive answer to any question worth asking (Norris & Ortega, 2006). Basic descriptive statistics go a long way toward answering our research questions, and they provide the basis for more sophisticated, synthetic answers as well.

Effect Sizes and CIs

A number of organizations, journals, and learned societies that have produced recommendations with regard to statistical reporting practices have insisted that researchers' emphasis should be shifted away from dichotomous *p* value interpretations of inferential statistical results and that the focus instead should be on effect sizes and CIs. It is dangerous, and simply incorrect, to believe that a *p* value tells you everything you need to know about a result (Larson-Hall, 2015a; Larson-Hall, 2010; Plonsky, 2013, 2014, in press; see also Norris, 2015). The effect of reporting a *p* value without an effect size is to encourage dichotomous thinking—either the result is significant or it is not. The reader

says, “Oh! The p value is less than .05, so the result is important” or “I see. . . the p value is greater than .05, so I can ignore the results.” However, p values depend on sample size and the amount of variability in the data and thus a small (statistically significant) p value may actually describe an unimportant result while a larger (statistically nonsignificant) one may hide an important result. Researchers are not seeing their results clearly if they rely on p values for interpretation.

The simplest effect size is one that is simply stated in the original measurement terms. For example, if two experimental groups were compared on their ability to speak fluently after different 2-month treatments, the most direct measure of effect size would be to say that Group A was able to produce on average 53 words per minute while Group B produced 44 words per minute, and we would see that Treatment A was more effective than Treatment B (assuming we had controlled any moderating variables). This answers the important question of how much effect our treatments had. A correlation is also an effect size, and we could report that the correlation between hours of language lab study and fluency was $r = .51$, meaning that about 25% ($.51^2$) of the variability in scores could be attributed to time spent in the language lab. A p value, on the other hand, is *not* an effect size, and cannot tell us anything about “how much” effect something we measured has had (Cumming, 2012).

The difference between means of different groups is a raw effect size, but many researchers also report standardized effect sizes so that effect sizes can be compared across different studies using different instruments or scales. Cohen’s d is a popular effect size to report for group differences, and it uses the mean as well as the standard deviations of the group to report a number that usually ranges between ± 3 (although it can be larger). Cohen’s d expresses the effect as a change in SD units and is obviously affected by the variability in group scores as well as the average score of the group (there are actually different ways to do this so one must be careful in choosing the standardizer and report how it was done). So, for example, when we reported the difference in group scores in the original measurement above, the difference in the original units was 9 words per minute, but we could report an effect size of Cohen’s $d = .50$, which would mean that the difference between groups was equal to $\frac{1}{2}$ of a standard deviation of the group scores. How important that difference might be will depend on various factors that the researcher will have to explain to the reader, but using a standardized effect size lets the reader compare effects across groups and also helps meta-analysts compare studies even if they use different original measures.

When effect sizes are not included in primary research reports, it is harmful for both consumers of primary research and future meta-analysis. Missing effect sizes limit readers' efforts to interpret the magnitude/strength of observed effects and encourage reliance on dichotomous p value interpretation. Unreported effect sizes also impede both *a priori* power analyses (to inform an appropriate sample size) and comparisons in future studies of similar relationships as well as inclusion of study effects at the meta-analytic level; they also introduce the potential for a biased set of effects when data reporting practices are associated with study outcomes (Plonsky, 2011). That is, when studies with larger effects report their data more thoroughly and are therefore more likely to be included in meta-analyses, meta-analytic results may overestimate actual effects in the population (see Plonsky, 2014).

For readers who are not familiar with CIs, we will provide a short explanation. A CI is a set of two numbers that represent a range of values where, with 95% confidence, you would expect some point estimate (e.g., a difference in mean scores between the groups) to appear if you repeated the same study again and again (Cumming, 2012). For a comparison of the difference between two means, if this range of scores goes through zero, you know it is possible that the true difference between groups is zero, which means that there is no statistically significant difference between them. But if the range does not cover zero, you can assume the difference between groups is statistically significant. That information is the same as what you could find from the p value of a t test, but CIs provide much more information. Besides telling us whether the difference between groups is statistical, they provide information about the confidence we can have in our single-point estimate of a mean difference between the groups (the smaller the interval, the less sampling error and the more confidence we have that we have nailed down a good estimate of the difference; the larger the interval, the more sampling error and the less confidence we have that we really have a good idea where the mean difference is) and the size of the effect (this is seen in how far away the smaller number is from zero, and how high it ranges; in the CI, you can interpret the numbers directly for the scale that you used, as opposed to a standardized effect size).

In order to illustrate how CIs can be interpreted, let's assume we have two groups: an experimental group who used an innovative piece of software that suggested appropriate vocabulary and grammar to students while writing papers, and a control group who simply wrote using standard methods. Our study measured the quality of the groups' final papers after two semesters of this procedure (based on a number of measures that added up to a total of 100 points) and our statistical test compared the groups to each other. Table 1 is an

Table 1 Examples of interpreting CIs

[4, 56.3]	There is a statistical difference between groups, since the difference in scores does not go through zero. However, the precision of the estimate is very poor. There might be as little as 4 points of difference between the groups in the population, or as much as 56.3 points (with 95% confidence, meaning that sometimes the difference might be even smaller or larger). Confidence intervals should help us avoid the dichotomous thinking that “there is a difference between groups” or “there is no difference between groups”. Instead, we see that the difference between the groups in the population is possibly very small (in this example, 4 points), or could be large (56.3 points). There is a lot of uncertainty here and without further research one could not know whether the approach was only slightly more useful than the standard approach, or wildly more useful. Further research should use bigger sample sizes to increase the reliability of the estimate, which would decrease the width of the confidence interval.
[-5.1, 8.3]	There is no statistical difference between groups, since the confidence interval crosses zero. Because the width of the confidence interval is rather small and precise, we could probably conclude that further research into the innovative treatment would not be very useful. In real life, it is not common to find such a small (precise) confidence interval.
[4.9, 10.6]	There is a statistical difference between groups. The difference for the population means could be as little as 4.9 points, or as much as 10.6 points, with 95% confidence. Because the CI is narrow, we could feel justified in feeling that we understood that the difference between groups is real, but the effect size is fairly small. The researcher would have to decide if a possible difference of 5 points improvement in the population were worth the extra possible time and expense involved in using the innovative software. If it does not take a lot of additional time to learn to use the software and if it is easily available, it is probably worth implementing (i.e., <i>practically</i> significant).
[-25.2, 0.3]	There is no statistical difference between groups, but it might be worth pursuing further research on the topic because the confidence interval only strays across zero for a very short distance. The difference between groups might be as small as zero, but it might be as large as 25.2 points (it does not matter that this is negative). Since the most likely population parameter is the one at the center of this interval (Kalinowski, 2014), if the CI were narrowed through further research it might be found to be exclude zero in its range, and this might be an effective procedure.

example of what kinds of CIs we might get for the difference in mean scores between the groups, and how we could interpret each one.

Notice that we have talked here about CIs by applying thinking about them to the difference in mean scores between two groups, but CIs can be applied to any other point estimate as well, such as a correlation coefficient (r), beta coefficients or R^2 estimates in a multiple regression, or eigenvalues in a factor analysis. There are some statistical techniques where complications arise in calculating CIs (Funder et al., 2013) or for which popular software does not compute them easily, but many researchers are working on ways to solve these problems and it should become increasingly easy to calculate CIs for any point estimate.²

The APA guidelines for reporting effect sizes and CIs indicate that they are extremely important to report. The APA manual states “For the reader to appreciate the magnitude or importance of a study’s findings, it is almost always necessary to include some measure of effect size . . . It can often be valuable to report an effect size not only in original units but also in some standardized or units-free unit (e.g., as a Cohen’s d value)” (p. 34). The APA also strongly recommends the reporting of CIs: “The inclusion of confidence intervals (for estimates of parameters, for functions of parameters such as differences in means, and for effect sizes) can be an extremely effective way of reporting results. . . . Confidence intervals combine information on location and precision . . . they are, in general, the best reporting strategy” (p. 34).

However, as shown in Figure 2, generally only a small portion of studies included in L2 meta-analyses to date have reported an effect size. Evidence on the reporting of CIs is even more disappointing, with their presence in meta-analyzed samples ranging from a mode of 0 (meaning 0 is the most common result) to just 5%. These findings, like those described in the previous paragraphs, indicate severe limitations to our ability to both consume and synthesize findings across primary studies. Perhaps even more disconcerting is the severe lack of synthetic-mindedness that these data represent. In other words, rather than considering each study to be contributing toward our collective understanding of a given set of relationships, primary authors appear to have very little concern for (a) practical (vs. statistical) significance, (b) cumulative knowledge, or for (c) the error present in every finding and the fact that no single study will provide a conclusive answer to the theoretical and practical questions posed in L2 research.

We have stated that effect sizes should be reported along with p values, and a number of journals in the L2 field now require that effect sizes be reported,

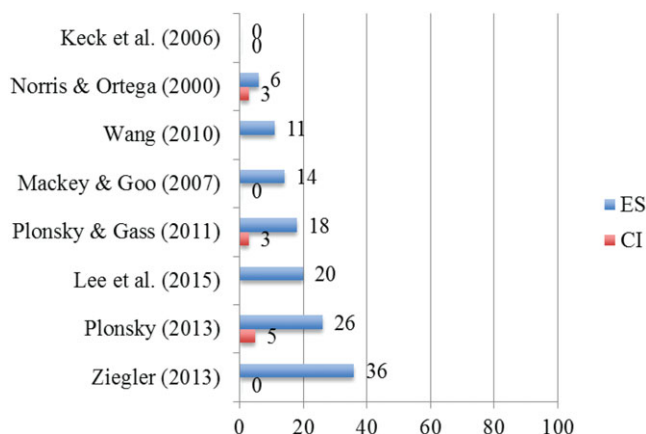


Figure 2 Reports from L2 meta-analyses that counted the number of primary studies reporting effect sizes and CIs (expressed as a % of the included sample).

so we feel confident that more and more researchers will provide effect sizes. However, we would also be happy to see researchers provide effect sizes and CIs, and leave out the p values altogether. Giving a confidence interval provides all the information of a p value, plus much more. Providing both a p value and a confidence interval is like getting two emails, one of which tells you you've won the lottery for 20 million dollars, which will be distributed over 30 years, so that after taxes you'll receive \$486,529 every year for 30 years, and another email that tells you you've won the lottery. The second email (the p value) is exciting news, but not nearly the whole story, and not necessary if you've already received the first e-mail.

The last point we would like to make in this section is that beyond providing effect sizes and CIs, meaningful *interpretation* should be given to the numbers. In most of these cases, authors (and reviewers/editors) treat the reporting of effect sizes as an end in itself rather than a means to more fully and informatively explain their findings (see Plonsky & Oswald, 2014). And when some interpretation is provided, it is almost always done so nearly as dichotomously as with p values. Authors need to report not only effect sizes but interpret them in the context of their area of research. Plonsky and Oswald (2014) give useful recommendations as to rough guidelines (and cautions) for initial interpretation of the magnitude of effect sizes (d values and correlation coefficients), based on the effect sizes observed in 346 primary studies and 91 meta-analyses. Their results show that Cohen's benchmarks of small ($d = .2$,

$r = .1$), medium ($d = .5$, $r = .3$), and large ($d = .8$, $r = .5$) underestimate effects typically found in L2 research. They recommend instead the following scales for roughly gauging small, medium, and large effects relative to others in the field, based on the 25th, 50th, and 75th percentiles for observed effects in their L2 research sample: for mean differences (d) resulting from between-groups contrasts, .4, .7, 1.0; for mean differences (d) resulting from within-groups/pre-post contrasts, .6, 1.0, 1.4; for correlation coefficients, .25, .4, .6. They also provide important additional considerations critical to providing maximally informative interpretations of effect sizes in L2 research.

Thus, saying that an effect is “large” is not enough. We urge authors to go beyond such benchmarks to add meaningful interpretation, given their knowledge of the relationships in question and the methods used to investigate them. This might include comparisons of effect sizes or CIs in similar studies as well as a consideration of the practical significance of the results. Here is an example of an interpretive statement added to an effect size: “The study found that the use of instructor-distributed candy in the classroom reduced language learning anxiety by $M = 4.3$ $SD = 2.1$ on a 50-point scale, Cohen’s $d = .53$. This is a somewhat small effect but given the low cost may be an approach that some instructors might want to implement.” In other words, the author of this study is arguing for beneficial practical results, given the tradeoff between the size of the effect and its cost for implementation. Obviously this is the author’s opinion but given the author’s expertise in the area, that is the person most likely to make the most informed evaluation. The readers have the original effect size and are also free to make their own evaluation of the interpretation for the study itself and by comparing its findings in other contexts.

Reliability

Reliability coefficients (Cohen’s kappa, Cronbach’s alpha, KR-20 estimates, inter-rater reliability, and so forth) should always be provided when any experimental measure is used. For example, if you administer a grammaticality judgment test in your experiment, you should report a measure of the instrument’s internal consistency. The same would apply for example, for a 20-item phonology test or a questionnaire that has results couched in numeric terms. Reliability is a measure of an instrument *with a certain sample of test takers*, so it should always be reported for each administration of a test. There is no *inherent* reliability of any test, and a report of the reliability of a test with a

different sample of test takers does not guarantee the reliability of a test for the current administration.

Omissions of reliability coefficients, as with *SDs*, also limit our ability to interpret study findings. For example, when no reliability coefficient is reported, it is very difficult to know whether small and/or nonstatistical effects are due to an instrument with low reliability or to some other factor such as a weak experimental manipulation, a truly small effect, an underpowered sample, or some other cause. What are the effects of low reliability? Lower reliability (internal consistency) would mean that a participant's scores vary more around their true ability. That is, performance or a score on each item, or a test on the whole, might be on average somewhat higher or lower than if it was a direct representation of what each participant actually knew in the L2 (their true score). When this value is correlated with another test of, say, aptitude (or any other variable), the resulting correlation is blurred and not as strong as it might be in reality. Low reliability might also produce larger *SDs* and *CI*s, leading to smaller *d* values (because the *SD* is in the denominator of the formula for *d*) as well as decreased likelihood of statistical findings. In essence low reliability means we cannot trust the results of the test, or any comparisons based on it, as much as we might hope we could.

Also at the primary level, missing reliability estimates limit future researchers in using the same or similar instruments for interpreting their data. At the secondary level, missing reliability estimates prevent meta-analysts from correcting or adjusting effect sizes for attenuation due to low reliability. The absence of such data also limits secondary researchers' ability to interpret the utility of different instrument types as well as their relationship to study outcomes.

Again, we stress that all reporting standards in the social and educational sciences that we are familiar with advocate reporting reliability estimates, including the APA. However, in the field of L2 research, they are often omitted. Figure 3 lists degree of reported reliability in all of the L2 meta-analyses for which these data were sought. As we show in Figure 3, the percentage of studies reporting reliability coefficients ranges from just 6% (L2 practice; Nekrasova & Becker, 2009) to 64% (L2 interaction; Plonsky & Gass, 2011) of the meta-analyzed samples (median = 41%). Much like we have seen with other types of data, an even smaller proportion of the studies in these samples went beyond simply reporting a reliability coefficient. That is, as with effect sizes and variance estimates, researchers seem to treat the reporting of reliability data as if it were an item to tick off a list of submission guidelines rather than a meaningful source of information and interpretive value. (For a synthesis of

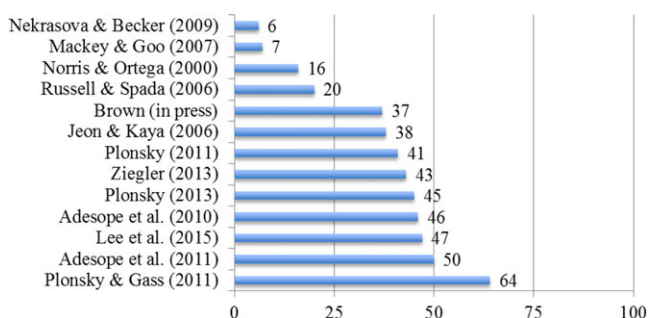


Figure 3 Reports from L2 meta-analyses that counted the number of studies reporting reliability (expressed as a % of the included sample).

reliability estimates and corresponding guidance on interpretations, see Plonsky & Derrick, 2015).

Before moving on we want to be clear about our perspective on the presence/absence of these data as a kind of proxy for transparency or even study quality. There is a critical distinction to be made between the reporting of instrument reliability or statistical power, for example, and *actual* or observed reliability or power. We are not equating the absence of a reliability coefficient or power analysis with low reliability or power. However, as we have explained throughout the chapter, these data are still necessary to allow for meaningful interpretation and evaluation of evidence at the primary and secondary/meta-analytic levels.

Visual Displays of Data

Most L2 researchers probably consider a graph to be a nice accompaniment to their research, but not an essential component. We would urge readers to consider graphs to be just as useful and necessary to both understanding and conveying their data findings in a published form as descriptive and inferential statistics. Edward Tufte (2001), a legend in statistical graphic design, has argued that they “can be more precise and revealing than conventional statistical computations” (p. 13). The APA has emphasized using visual displays of data as a way of understanding data even before any inferential procedures are conducted. A recommendation we agree with wholeheartedly is this: “As soon as you have collected your data, before you compute *any* statistics, *look at your data*. . . . If you assess hypotheses without examining your data, you risk publishing nonsense” (Wilkinson & Task Force on Statistical Inference,

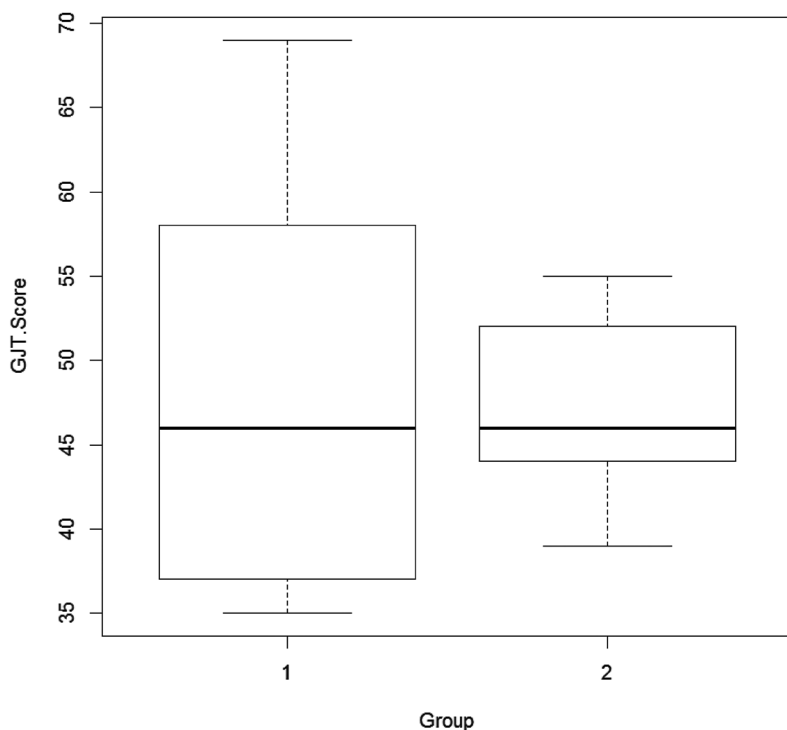


Figure 4 A boxplot for two experimental groups.

1999, p. 597, *italics in the original*). Wilkinson and the Task Force on Statistical Inference (1999) also recommend using graphics to diagnose possible problems with statistical assumptions like normality and homogeneity of variances, noting that numeric summaries for such assumptions are often overly sensitive and not accurate. For reporting results, Wilkinson and the Task Force on Statistical Inference (1999) recommend providing both tables and figures to cater to individual preferences of readers. These recommendations lead us to also urge L2 researchers to make broad use of figures in both understanding results before any writing takes place as well as in research reports themselves.

Here is an example of how examining data graphically might help save embarrassment later. Say you have performed an experiment to see whether a computer-administered treatment improves scores on a grammaticality judgment test and you have two groups (a control and a treatment). You look at a boxplot of scores for the two groups on the test (Figure 4). The boxplot shows that the groups have the same median score, although there was more variation

in scores in Group 1 than in Group 2 because the box and the whiskers are wider. Next, say you run a t test and you see that the difference results in a p value of .002, which you would want to interpret as showing there was a statistical difference between groups. However, the graphics would argue against finding a statistical difference between groups, and as you go back you realize that instead of looking at the results of the t test, you actually were looking at the results of the Levene's test for equality of variances in the print out! The low p value for Levene's test indicates that variances are not equal (which could also be seen in the boxplot), but now you see that the appropriate p value for the t test is $p = .70$ (and you notice that there is a 95% CI of $[-48, 67]$, indicating that the difference between the groups goes through zero and is extremely wide). Obviously, the entire tenor of your article will change depending on whether you think there is a significant difference between the groups' scores or not, and using graphics helped you align the reality of the data distribution with your inferential statistics.

Another example, also hypothetical, concerns a critical period study where you have gathered information about the age of acquisition of an L2 (AoA), current chronological age (AGE), and length of residence (LOR). A scatterplot matrix that shows all three variables (Figure 5) crossed by all of the other variables quickly reflects a number of anomalies. For one thing, the distribution of age (by itself; see the topmost box) shows a fairly normal distribution except for a number of cases at age 99. This most likely represents a practice where the researcher used the number 99 to code missing cases, but this would be dangerous in the case of age, and the graphic shows that the researcher would need to exclude this particular value from the age data. Further, the scatterplot of age with the variable of LOR, found in the third row down, first column of the graphic, shows that some participants have an LOR longer than their age (the 3 points found to the upper left-hand side of the mass of data points represent this impossible situation). The researcher would need to check their data and eliminate these values from further statistical consideration. Thus we see how easy it is to spot anomalies by using graphics.

In L2 research, visual displays of data appear in around half (or less) of the studies (as observed in meta-analyses: 53% in Plonsky, 2013; 34% in Nekrasova & Becker, 2009; and 56% in Norris & Ortega, 2000). A survey of three L2 journals (*The Modern Language Journal* [MLJ], *Language Learning* [LL], and *Studies in Second Language Acquisition* [SSLA]) found that over the entire publication history of these journals the percentage of articles with graphics was fairly low (24%, 34%, and 48%, respectively; see Larson-Hall, 2015b). A

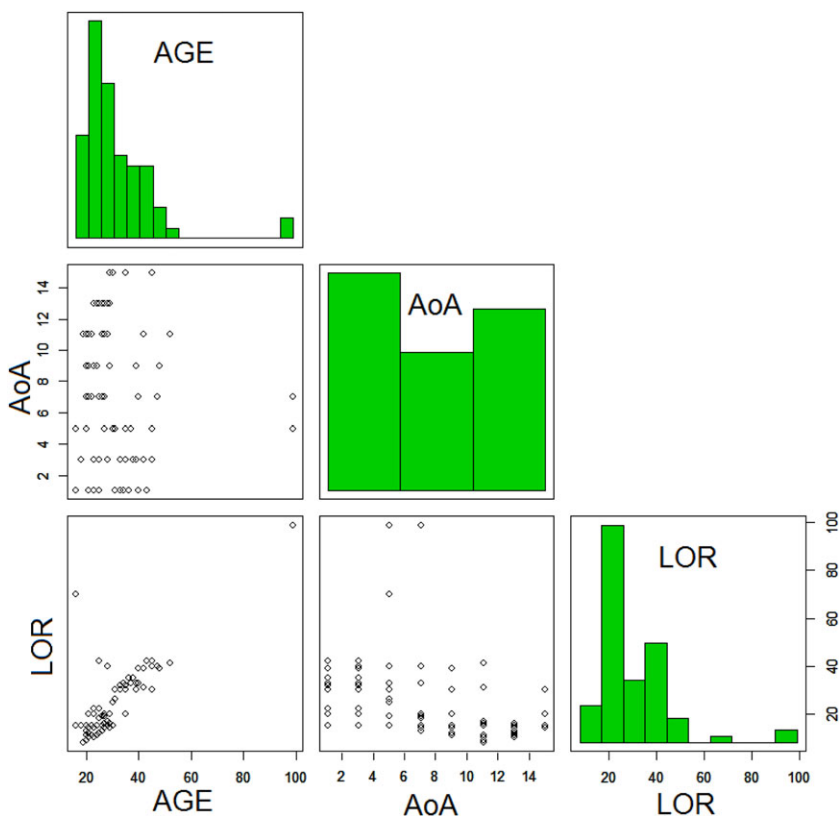


Figure 5 A scatterplot matrix of data that show anomalies.

lack of visuals indicates a missed opportunity to convey findings succinctly to readers.

However, as with effect sizes, *SDs*, and *CI*s, just providing a graphic is not enough. The graphic should be appropriate for the data, show the data in a way that lets the reader see individual phenomena as well as patterns in the data when possible, and should be interpreted for the reader as well (Hudson, in press; Larson-Hall, 2015b). Most graphics published in the L2 field are, we argue, data-poor and not very informative. Larson-Hall (2015b) found that over the entire publication history of the three L2 journals mentioned above, in those articles that did publish graphics, the percentage of articles using barplots and/or line graphs was 79% (*MLJ*), 70% (*LL*), and 79% (*SSLA*). Most of the features of graphics that are useful, such as the ability to provide visual insights

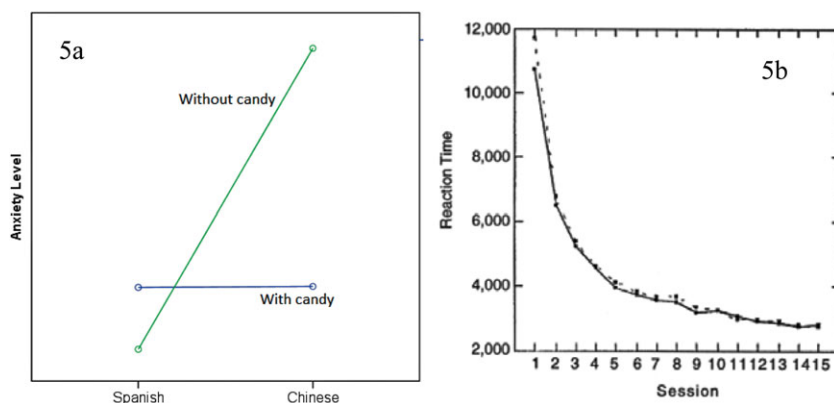


Figure 6 Examples of line graphs (6A is based on fabricated data, 6B is from DeKeyser, 1997).

that cannot be readily discerned from numbers only, or the ability to show data that could provide information supporting statistical inferences or to highlight numerical anomalies, cannot usually be taken from barplots and line graphs because they do not provide any information about the variability in the data nor how individual data points are distributed and thus do not provide a level of detail that lets the eye make visual comparisons.

The APA task force report on statistical inference suggests that graphics such as barplots and line graphs lack essential statistical information, and beyond this, the omission of data about the shape of the distribution actually “*hinders scientific evaluation*” (Wilkinson & Task Force on Statistical Inference, 1999, p. 601, emphasis added). Line graphs might provide a lot of visual impact when showing an interaction of effects, as in Figure 6, but one wonders if the space they take up is worth the few data points that they show (only *four* points in Figure 6a). When line graphs chart a series of measurements over time that show some kind of progression or pattern as in Figure 6b, they are more useful and informative.

As for the barplot, Tufte (2014) says, “[t]he bar chart wastes space; you could show at least 100 numbers in the space that now shows 1 number.” A barplot can be used appropriately when the data consist of counts, such as the information found in Figures 1, 2, and 3 of this paper. Larson-Hall and Herrington (2009) have argued, using L2 data to illustrate the point, that in the case of mean values the boxplot is a better option than the barplot and

can reveal more about the distribution of the data. Even better, a boxplot with overlaid dots representing individual scores shows group patterns, the shape of the distribution of the data, and also the uncertainty in the data. It is a good choice for a data-accountable graphic for group data (see Larson-Hall, 2015a, Chapter 8 for instructions on how to create this type of graphic using the R statistical program and the `ggplot2` package.). Compare the amount of data that is provided in a barplot (Figure 7A) and a boxplot with overlaid dots for the same data (Figure 7B, taken from Larson-Hall, 2008). The barplot with overlaid dots provides all of the data in front of the reader's eyes, enabling them to get the overall picture and to judge the distribution of the data. Note that the Earlier starters have a higher median score and somewhat smaller variance for the 25th–50th percentile (enclosed by the grey box) than the Later starters, but also see that the entire range is nearly the same length, and that the Earlier category has a couple of outliers. In other words, you can see the pattern of the data as well as the uncertainty (the variation) in it.

Another type of graphic that works well with repeated measures designs is the parallel coordinate plot. Figure 8 shows such a plot using data from Lyster (2004) with a pretest, posttest, and delayed posttest. There are four different groups, each of which received a different treatment. The individual lines show scores on a written test. The higher each line is, the higher the score. The darker black line superimposes the mean score on the graphic.³ The FFI-prompt group statistically outperformed all of the other groups on the immediate and delayed posttest, but the use of a data-accountable graphic that shows the individual data as well as overall trends lets the reader see that not every individual in that group showed marked improvement (although a much larger number *did* show marked improvement in the FFI-prompt group than in any of the other groups).

The scatterplot, which is more well-known, is also an excellent data-accountable graphic that provides all of the individual data for the viewer to see and to judge just how good the fit is to a straight regression line and a local fit line (we recommend the Loess line, which follows the pattern of the data). Again, the viewer can see all of the variability and uncertainty in the data before their eyes, and thus has a rubric against which to judge the author's interpretation of the data.

In summary, we advise readers to use graphics in their research reports, but to do so with an eye toward providing the most informative, data-rich and appropriate graphic they can. Such graphics will provide the most information for readers and let the data speak for itself.

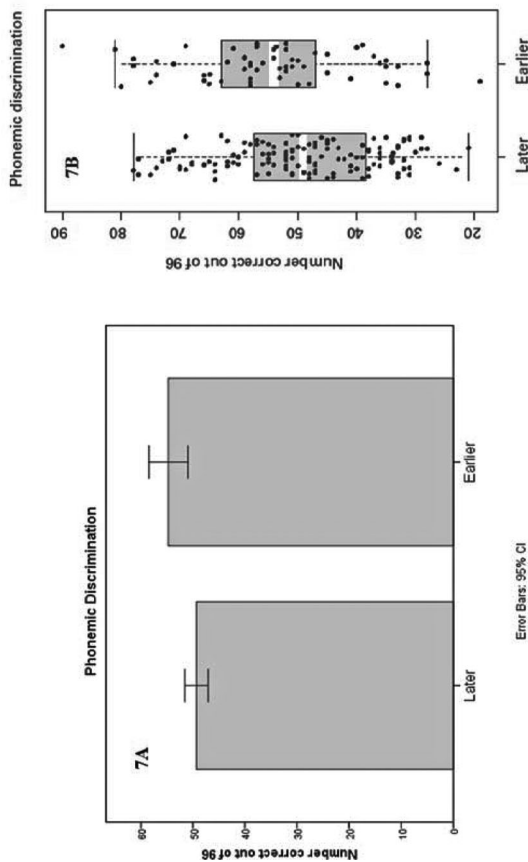


Figure 7 Examples of a barplot and a boxplot with overlaid dots using the same data (from Larson-Hall, 2008).

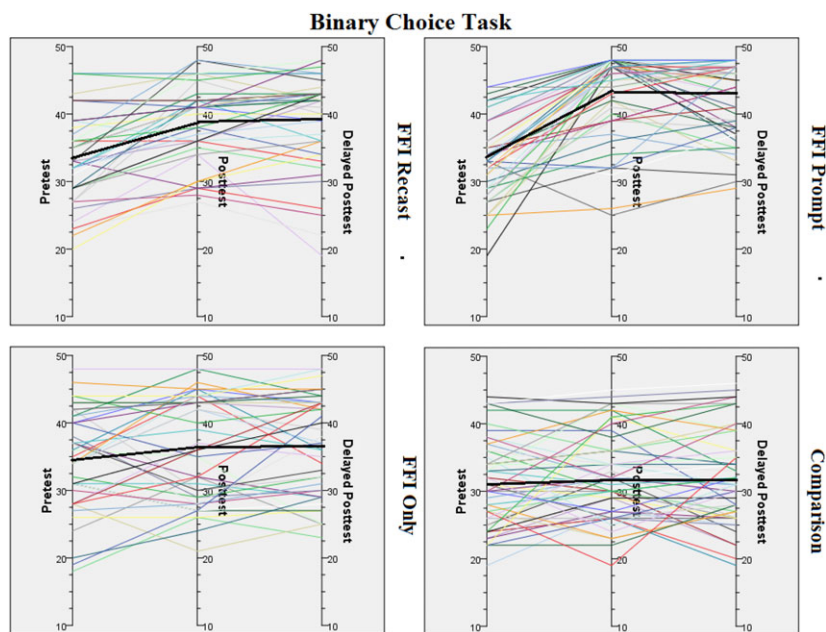


Figure 8 Parallel coordinate plot in the binary choice task from Lyster (2004) data, where 3 groups received form-focused input (FFI) on the gender of articles in French plus some type of correction (none, recast, or prompt), and one group received no special instruction (Comparison).

Data Sharing

Finally, what is often not reported is the basis for the previously mentioned features, the raw data. As noted previously, APA guidelines for the use of the data upon which empirical articles are based states that researchers should share their data with other researchers for the sake of reanalysis. The APA notes that data form the foundation of our science, but that since psychologists are dealing with data from humans and must protect their identities, data sharing can be more complicated than in fields such as biology or economics. This is of course true for much research in the L2 field as well. However, APA Ethical Standard 8.14 “stipulates that psychologists do not withhold their data from other competent professionals who seek to verify substantive claims” (Breckler, 2009, para. 7). The Linguistic Society of America also made a recent statement (Good, 2010) urging linguists to “make the full data sets behind publications available, subject to all relevant ethical and legal concerns” (para. 6)

and also “when serving as reviewers, expect full data sets to be published (again subject to legal and ethical considerations) and expect claims to be tested against relevant publicly available datasets” (para. 8). The APA has also recommended new guidelines in this regard, which the journal *Psychological Science* is following. Regarding data sharing and openness, this journal will award badges for providing a “URL, doi, or other permanent path for accessing the data in a public, open access repository” (“Badges to acknowledge,” 2013).

We agree with these task forces and urge researchers to follow more open practices, which would allow other researchers in the future to verify statistical claims that are based upon the data, as well as perform different analyses of the data as new information comes to light, and possibly combine databases for stronger inferences and more insight into various phenomena. Researchers should be willing to share data unless there are some compelling reasons of privacy or legal conflicts that make it impossible.

A recent study by Vines et al. (2014) showed that it is likely that without external methods of storage, most data that support research studies will become lost. In this study the researchers tried to contact the authors of 516 papers published 2–20 years prior to obtain their data and were only able to obtain data from 20% of the studies. The main obstacles were obsolete emails and broken data storage devices. Likewise in the L2 domain, of the 255 authors contacted by Plonsky et al. (in press), only 36 (14%) replied with the requested datasets. Because it is expensive and time-consuming to obtain the data for research, it seems reasonable to want to, as a field, provide a place where researchers could access data from previously published reports. Until some agreed-upon space becomes available for data storage, we urge researchers to publish the raw data from their own research. Current internet-based publishing allows space for data to be published in online appendices that need not add to the length of articles. Another option is to make datasets available through individual researchers’ Web sites. At the very least, journal reviewers should expect to see raw data submitted (even if it cannot be published) and should use it to verify statistical claims.

Summary

In this paper we have attempted to provide information about the current standards of reporting for quantitative research, drawing primarily on the recommendations of the APA. Although our findings show that there are some deficits within L2 quantitative research in reporting in all of the areas we have surveyed, namely descriptive statistics, effect sizes and CIs, reliability

measurements, graphics, and data sharing, we also feel optimistic that changes are taking place, especially as journals have begun to recognize the need for top-down methodological reforms and standards for quantitative reporting.

It seems that we may have reached what Heidi Byrnes, editor of *MLJ*, called in an editor's note (2013) a "methodological turn" (p. 825), a point where researchers may finally be paying attention to the recommendations of statistically minded practitioners of various social sciences and, consequently, reflecting more on methodological issues in our own field. Two studies—Plonsky (2014) and Plonsky and Gass (2011)—have focused specifically on the dynamic nature of L2 reporting practices in the field, and the evidence they present is encouraging. Both studies found that nearly all of the data types included in our review above are improving appreciably over time. Plonsky (2014), for example, found the percentage of studies reporting a mean without an *SD* decreased from 41% to 24% of the sample from the 1990s and 2000s, respectively. Both studies also found increases, some dramatic, in the reporting of effect sizes (3–42% and 0–27% in the 1990s and 2000s in Plonsky, 2014, and Plonsky & Gass, 2011, respectively), CIs (0.4–7% and 0–4% in Plonsky, 2014), reliability coefficients (38–50% and 41–76%), research questions (70–87% in Plonsky, 2014), a priori alpha levels (16–26% in Plonsky, 2014), testing of statistical assumptions (13–19% and 0–5%), and power analyses (0–2% and 0–3%). As can be seen, although there have been improvements, there are still major strides to be made, especially in areas such as the use of CIs, and we must work toward further improvements in data reporting practices.

To that end, we will now reiterate and briefly outline an initial set of guidelines for reporting practices in quantitative L2 research. Our recommendations are based largely on the deficiencies described above as well as existing guidelines described in each section. Rather than presenting our recommendations as definitive, we hope to see further refinement over time as awareness of the importance of transparency grows among L2 researchers. We believe that ideal or even acceptable data reporting practices are not necessarily forthcoming without a major shift in the way this issue is addressed by journals, learned societies such as the American Association for Applied Linguistics (AAAL), and graduate training such that transparency becomes integral to the culture of quantitative L2 research (see Plonsky, 2014). For example, Ellis (2000) established for the first time in the L2 domain the requirement that articles in *LL* report effect sizes. Since then, at least three other L2 journals have released similar policies on the reporting of effect sizes: *Language Learning & Technology*, *MLJ*, and *TESOL Quarterly* (*TQ*). Notably, the new (as of late 2013) guidelines on the website of *MLJ* also stress the importance of meaningful

interpretation of effect sizes beyond a generic label such as “small,” “medium,” or “large” (see Plonsky & Oswald, 2014). *TQ* and *MLJ* have also enacted additional guidelines wherein authors are required or at least encouraged to include CIs. However, our intent is not to just persuade authors to include additional statistics in their reports, but also to persuade them to take on a synthetic mindset and to view study effects in relative terms as opposed to dichotomously (i.e., significant/nonsignificant).

Descriptive Statistics

1. Report descriptive statistics thoroughly and consistently. Provide:

- A measure of the average trend of the data for all configurations of the data that are analyzed (the mean or more rarely, the median)
- A measure of the variability in the data (the standard deviation or a confidence interval for the mean and both if possible would be best; if a median is used then interquartile ranges)
- Research questions and/or hypotheses
- An a priori alpha level and exact p values (when inferential statistics are used; these are not necessary if confidence levels are provided)
- Information about whether statistical assumptions were checked (and, if unmet, any accommodations made)
- Whether statistical power or precision (if CIs are being used) was considered in the study's design in order to determine an appropriate sample size

Treat means and standard deviations as a meaningful source of information and interpretative value. Seek to comment on them relative to other values in your paper as well as to values in similar work, thinking about your work as one piece of a larger whole.

Effect Sizes

2. Provide effect sizes for all configurations of the data that are analyzed.

Effect sizes can be given either in the original measurement or as standardized effect sizes (common ones include Cohen's d for group data or the correlation coefficient r for correlational data) or ideally both. Interpret effect sizes not only by calling them small, medium, or large but compare them to other effect sizes in your own study and in other similar studies. Point out whether the effect sizes indicate a practical effect, one that other researchers might consider important, given the time, effort, money, or other resources needed for implementation. Effect sizes can be explained in light of many factors

including study design, the sample, and the relationships in question (see Plonsky & Oswald, 2014).

3. Provide CIs for all configurations of the data that are analyzed.

At the time this is not always possible (calculations for CIs are not always available in some situations), but insofar as it is possible, CIs are desirable and may replace *p* values from statistical tests. Interpret CIs by referring to their width (precise or not) and distance away from zero (their effect size). Intervals that pass through zero do not always indicate a lack of effect and intervals that avoid zero do not always indicate a practically important effect. Further interpret CIs by comparing them with other intervals in your own study and across other studies.

Reliability Measures

4. Give reliability coefficients such as Cohen's kappa, Cronbach's alpha, KR-20 estimates, or inter-rater reliability for every measure that is used.

Reliability estimates should be given for every administration of a test, regardless of whether it is a standardized test that has been used many times before (such as the TOEFL) or a test you have created yourself. Information about the reliability of an instrument, whether internal consistency or inter-rater, is critical to evaluating the data in a study and therefore a study's findings. Future studies and replications in particular also benefit greatly from reliability estimates. Treat this number as a meaningful source of information and interpretative value and compare it to other administrations of the same or similar tests.

Visual Displays of Data

5. Before you compute any statistics, look at your data visually.

Check for anomalies. For group data, boxplots can be helpful. For correlational data, multiple scatterplots are useful.

6. Use graphics to check assumptions for statistical tests, such as the assumptions of normal distribution or homogeneity of variances.

Graphics are often a more reliable source for understanding the distribution of your data than numeric summaries such as Levene's test.

7. Include data-rich informative graphics for your main analyses.

For group data, boxplots with overlaid dots show individual data as well as group trends. For repeated (group) data, the parallel coordinate plot also shows individual data and can be supplemented with a line showing the mean value of the groups at the different times. For correlational data or multiple regression, scatterplots with overlaid Loess lines (and additionally regression lines, if desired) again show all of the data being considered but also show the bigger picture and trends as well.

Data Sharing

8. Ideally, publish the raw data upon which the quantitative conclusions of your study are based, either by including it as an appendix to your submitted paper, or archiving it in a database.

If ethical or legal concerns make this unfeasible, at least include the raw data for reviewers to check the statistical reasoning behind your study, and share the data with other competent professionals who seek to check your quantitative calculations. If publishing the raw data is not possible because of space considerations, seek to find an online repository to make the data available.

Meta-Analysis

9. Think and interpret all findings synthetically.

When you report your results, keep in mind that your work is part of a larger body of research and you want to make sure your work can be included in a synthesis or meta-analysis of your subdiscipline sometime in the future. This involves not only thorough reporting of descriptive statistics (including effect sizes and CIs) and exact test statistics (e.g., t , F , χ^2); synthetic thinking at the primary level also involves understanding that one study is part of a particular line of inquiry that cannot provide the one definitive answer, but will contribute toward advancing understanding as part of a body of knowledge.

Final revised version accepted 15 February 2015

Notes

- 1 Focus on Form (FonF) is a focus on forms within the context of meaningful interaction, as opposed to Focus on Forms (FonFS), which is explicit focus on grammatical forms only.

- 2 The first author's second edition of her book *A guide to doing statistics in second language research using SPSS* (& now R too) provides help in using statistical software to calculate confidence intervals on all point estimates for the basic statistical tests included in the textbook.
- 3 This graphic was created with SPSS but that program does not superimpose mean lines on top of the individual lines. The lines were added using the Paint program and so they are necessarily not as precise as if a computer program had added them. Thanks to John Norris for the suggestion.

References

- Abraham, L. B. (2008). Computer-mediated glosses in second language reading comprehension and vocabulary learning: A meta-analysis. *Computer Assisted Language Learning*, 21, 199–226.
- Adesope, O. O., Lavin, T., Thompson, T., & Ungerleider, C. (2010). A systematic review and meta-analysis of the cognitive correlates of bilingualism. *Review of Educational Research*, 80, 207–245.
- Adesope, O. O., Lavin, T., Thompson, T., & Ungerleider, C. (2011). Best practices in teaching literacy to ESL immigrant students: A meta-analysis. *British Journal of Educational Psychology*, 81, 629–653.
- American Educational Research Association. (2006). Standards for reporting on empirical social science research in AERA publications. *Educational Researcher*, 35(6), 33–40.
- American Psychological Association. (2010). *Publication manual of the American Psychological Association* (6th ed.). Washington, DC: Author.
- APA Publications and Communications Board Working Group on Journal Article Reporting Standards. (December 2008). *American Psychologist*, 63(9), 839–851.
- Badges to acknowledge open practices. (2013, February 20). *Open Science Framework Beta*. Retrieved January 11, 2014, from <https://osf.io/tvyxz/wiki/aandi>
- Biber, D., Nekrasova, T., & Horn, B. (2011). The effectiveness of feedback for L1-English and L2 writing development: A meta-analysis. *TOEFL iBT Re-search Report No. TOEFLiBT-14*. Princeton, NJ: Educational Testing Service.
- Bowles, M. (2010). Features that make a task amenable to think-aloud: A meta-analysis of studies investigating the validity of think-alouds on verbal tasks. In *The think-aloud controversy in second language research* (Chap. 3). New York: Routledge.
- Breckler, S. (2009). Dealing with data. *American Psychological Association: Science Directions*. Retrieved February 10, 2014, from <http://www.apa.org/monitor/2009/02/sd.aspx>
- Brown, D. (in press). The type and linguistic foci of oral corrective feedback in the L2 classroom: A meta-analysis. *Language Teaching Research*.
- Byrnes, H. (2013). Notes from the editor. *Modern Language Journal*, 97, 825–827.

- Catford, J.C. (1998). *Language Learning* and the applied linguistics: A historical sketch. *Language Learning*, 48(4), 465–496.
- Chaudron, C. (2001). Progress in language classroom research: Evidence from *The Modern Language Journal*, 1916–2000. *The Modern Language Journal*, 85, 57–76.
- Cooper, H. (2011). *Reporting research in psychology: How to meet journal article reporting standards*. Washington, DC: American Psychological Association.
- 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: Routledge.
- DeKeyser, R. (1997). Beyond explicit rule learning: Automatizing second language morphosyntax. *Studies in Second Language Acquisition*, 19, 195–221.
- Dinsmore, T. H. (2006). Principles, parameters, and SLA: A retrospective meta-analytic investigation into adult L2 learners' access to Universal Grammar. In J. M. Norris & L. Ortega (Eds.), *Synthesizing research on language learning and teaching* (pp. 53–90). Philadelphia: John Benjamins.
- Ellis, N. C. (2000). Editorial statement. *Language Learning*, 50, xi–xiii.
- Funder, D., Levin, J., Mackie, D., Morf, C., Vazire, S., & West, S. (2013). Improving the dependability of research in personality and social psychology: Recommendations for research and educational practices. *Personality and Social Psychology Review*, 18(1), 3–12.
- Goldschneider, J. M., & DeKeyser, R. M. (2001). Explaining the natural order of L2 morpheme acquisition: A meta-analysis of multiple determinants. *Language Learning*, 51, 1–50.
- Good, J. (2010, January 11). Linguistics society debates data sharing. *Open Access News*. Retrieved February 10, 2014, from <http://legacy.earlham.edu/~peters/fos/2010/01/linguistics-society-debates-data.html>.
- Grgurović, M., Chappelle, C. A., & Shelley, M. (2013). A meta-analysis of effectiveness studies on computer technology-supported language learning. *ReCALL*, 25, 165–198.
- Henning, G. (1986). Quantitative methods in language acquisition research. *TESOL Quarterly*, 20, 701–708.
- Hudson, T. (in press). Presenting quantitative data visually. In L. Plonsky (Ed.), *Advancing quantitative methods in second language research*. New York: Routledge.
- Jeon, E. H., & Kaya, T. (2006). Effects of L2 instruction on interlanguage pragmatic development: A meta-analysis. In J. M. Norris & L. Ortega (Eds.), *Synthesizing research on language learning and teaching* (pp. 165–211). Philadelphia: John Benjamins.
- Kalinowski, P. (2014). Understanding confidence intervals (CIs) and effect size estimation. Retrieved February 6, 2014, from <http://www.psychologicalscience.org/index.php/publications/observer/2010/april-10/understanding-confidence-intervals-cis-and-effect-size-estimation.html>

- Keck, C. M., Iberri-Shea, G., Tracy-Ventura, N., & Wa-Mbaleka, S. (2006). Investigating the empirical link between task-based interaction and acquisition: A meta-analysis. In J. M. Norris & L. Ortega (Eds.), *Synthesizing research on language learning and teaching* (pp. 91–131). Philadelphia: John Benjamins.
- Kieffer, M. T., Lesaux, N. K., Rivera, M., & Francis, D. J. (2009). Accommodations for English language learners taking large-scale assessments: A meta-analysis on effectiveness and validity. *Review of Educational Research*, 79, 1168–1201.
- Larson-Hall, J. (2008). Weighing the benefits of studying a foreign language at a younger starting age in a minimal input situation. *Second Language Research*, 24, 35–63.
- Larson-Hall, J. (2010). *A guide to doing statistics in second language research using SPSS*. New York: Routledge.
- Larson-Hall, J. (2015a). *A guide to doing statistics in second language research using SPSS and R*. New York: Routledge.
- Larson-Hall, J. (2015b). Moving beyond the bar plot and line graph to create informative and attractive graphics. Manuscript in preparation.
- Larson-Hall, J., & Herrington, R. (2009). Improving data analysis in second language acquisition by utilizing modern developments in applied statistics. *Applied Linguistics*, 31, 368–390.
- Lee, S.-K., & Huang, H.-T. (2008). Visual input enhancement and grammar learning: A meta-analytic review. *Studies in Second Language Acquisition*, 30, 307–331.
- Lee, J., Jang, J., & Plonsky, L. (in press). The effectiveness of second language pronunciation instruction: A meta-analysis. *Applied Linguistics*.
- Li, S. (2010). The effectiveness of corrective feedback in SLA: A meta-analysis. *Language Learning*, 60, 309–365.
- Lin, W.-C., Huang, H.-T., & Liou, H.-C. (2013). The effects of text-based SCMC on SLA: A meta-analysis. *Language Learning & Technology*, 17, 123–142.
- Loewen, S., & Gass, S. (2009). Research timeline: The use of statistics in L2 acquisition research. *Language Teaching*, 42(2), 181–196.
- Lyster, R. (2004). Differential effects of prompts and recasts in form-focused instruction. *Studies in Second Language Acquisition*, 26, 399–432.
- Mackey, A., & Goo, J. (2007). Interaction research in SLA: A meta-analysis and research synthesis. In A. Mackey (Ed.), *Conversational interaction in second language acquisition: A collection of empirical studies* (pp. 407–451). New York: Oxford University Press.
- Nekrasova, T., & Becker, T. (2009). *Effectiveness of practice: A research synthesis and quantitative meta-analysis*. Unpublished manuscript.
- Nicol, A. A. M., & Pexman, P. M. (2010a). *Presenting your findings: A practical guide for creating tables*. Washington, DC: American Psychological Association.
- Nicol, A. A. M., & Pexman, P. M. (2010b). *Displaying your findings: A practical guide for creating figures, posters, and presentations*. Washington, DC: American Psychological Association.

- Norris, J. M. (2015). Statistical significance testing in second language research: Basic problems and suggestions for reform. *Language Learning*, 65(S1), 97–126.
- Norris, J. M., & Ortega, L. (2000). Effectiveness of L2 instruction: A research synthesis and quantitative meta-analysis. *Language Learning*, 50, 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). Philadelphia: John Benjamins.
- Nunan, D. (1991). Methods in second language classroom-oriented research: A critical review. *Studies in Second Language Acquisition*, 13, 249–274.
- Pennock-Roman, M., & Rivera, C. (2011). Mean effects of test accommodations for ELLs and non-ELLs: A meta-analysis of experimental studies. *Educational Measurement: Issues and Practice*, 30, 10–28.
- Phakiti, A. (2010). Analysing quantitative data. In B. Paltridge & A. Phakiti (Eds.), *Continuum companion to research methods in applied linguistics* (pp. 39–49). London: Continuum.
- Plonsky, L. (2011). The effectiveness of second language strategy instruction: A meta-analysis. *Language Learning*, 61, 993–1038.
- 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., & Derrick, D. J. (2015). A meta-analysis of reliability coefficients in second language research. Manuscript submitted for publication.
- Plonsky, L., Egbert, J., & LaFlair, G. T. (in press). Bootstrapping in applied linguistics: Assessing its potential using shared data. *Applied Linguistics*.
- 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.
- Russell, J., & Spada, N. (2006). The effectiveness of corrective feedback for the acquisition of L2 grammar: A meta-analysis of the research. In J. M. Norris & L. Ortega (Eds.), *Synthesizing research on language learning and teaching* (pp. 133–164). Philadelphia: John Benjamins.
- Tufte, E. (2001). *The visual display of quantitative information* (2nd ed.). Cheshire, CT: Graphics Press.
- Tufte, E. (2014, January 16). Sparkline theory and practice Edward Tufte [Online forum comment]. Retrieved from http://www.edwardtufte.com/bboard/q-and-a-fetch-msg?msg_id=0001OR

- Vines, T., Albert, A., Andrew, R., Débarre, F., Bock, D., Franklin, M., et al. (2014). The availability of research data declines rapidly with article age. *Current Biology*, 24(1), 94–97.
- Wainer, H. (1996). Depicting error. *American Statistician*, 50, 101–111.
- Wa-Mbaleka, S. (2006). *A meta-analysis investigating the effects of reading on second language vocabulary learning*. Unpublished doctoral dissertation, Northern Arizona University, Flagstaff, AZ.
- Wang, L.-J. (2010). *A meta-analysis of empirical studies on the effects of extensive reading*. Unpublished master's thesis, National Tsing Hua University, Taiwan.
- Wilkinson, L., & Task Force on Statistical Inference. (1999). *Statistical methods in psychology journals: Guidelines and explanations*. *American Psychologist*, 54, 594–604. Available from tinyurl.com/tfsi1999
- Wu, C. (1991). *A meta-analysis of intensive audio-lingual language programs: A meta-analysis of intensive audio-lingual language programs: Implications for English as a foreign language instruction in China*. Unpublished doctoral dissertation, University of San Francisco.
- Ziegler, N. (2013). *Synchronous computer-mediated communication and interaction: A research synthesis and meta-analysis*. Unpublished doctoral dissertation, Georgetown University.