LANGUAGE LEARNING
A Journal of Research in Language Studies

Language Learning ISSN 0023-8333

METHODOLOGICAL REVIEW ARTICLE

How Big Is "Big"? Interpreting Effect Sizes in L2 Research

Luke Plonsky[a] and Frederick L. Oswald[b]

[a]Northern Arizona University and [b]Rice University

The calculation and use of effect sizes—such as *d* for mean differences and *r* for correlations—has increased dramatically in second language (L2) research in the last decade. Interpretations of these effects, however, have been rare and, when present, have largely defaulted to Cohen's levels of small ($d = .2, r = .1$), medium (.5, .3), and large (.8, .5), which were never intended as prescriptions but rather as a general guide. As Cohen himself and many others have argued, effect sizes are best understood when interpreted within a particular discipline or domain. This article seeks to promote more informed and field-specific interpretations of *d* and *r* by presenting a description of L2 effects from 346 primary studies and 91 meta-analyses ($N >$ 604,000). Results reveal that Cohen's benchmarks generally underestimate the effects obtained in L2 research. Based on our analysis, we propose a field-specific scale for interpreting effect sizes, and we outline eight key considerations for gauging relative magnitude and practical significance in primary and secondary studies, such as theoretical maturity in the domain, the degree of experimental manipulation, and the presence of publication bias.

Keywords effect sizes; quantitative research methods; meta-analysis; practical significance

Introduction

The introduction of meta-analysis in applied linguistics has led to an increased awareness and modest reforms of research and reporting practices in the field (Plonsky, 2014). For instance, more thorough reporting of descriptive statistics has been observed along with interest in and consideration of practical significance expressed by effect sizes (Plonsky & Gass, 2011). However, very little is known about how to interpret effect sizes beyond Cohen's (1988) general

Correspondence concerning this article should be addressed to Luke Plonsky, Northern Arizona University, Department of English, P.O. Box 6032, Flagstaff, AZ 86011. E-mail: luke.plonsky@nau.edu

DOI: 10.1111/lang.12079

benchmarks of $d = .2$, $r = .1$ (small); .5, .3 (medium); and .8, .5 (large). This article comprises two parts that reflect our two goals with this research. In Part I, we provide an empirically based starting point for informing the numerical interpretation of effect sizes in second language (L2) research, and in Part II we examine in some depth a number of broader considerations that should inform L2 researchers' interpretations of effect sizes. We begin, though, by echoing previous calls for the importance and usefulness of effect sizes in L2 research (e.g., Norris & Ortega, 2006; Plonsky, 2012a).

The Case for Effect Sizes

The most compelling case for effect sizes, defined as a "quantitative reflection of the magnitude of some phenomenon ... of interest" (Kelley & Preacher, 2012, p. 140), is that they are indices of practical significance, which supplements information from traditional statistical significance tests. The academic controversy surrounding statistical significance, where practical significance is often offered as an important alternative or supplement, dates back to the first half of the 20th century (see Cohen, 1994; Thompson, 2001) and includes occasional critiques of *p* values within L2 research (Larson-Hall & Plonsky, in press; Nassaji, 2012; Norris & Ortega, 2006; Oswald & Plonsky, 2010; Plonsky, 2009, 2011a). Nevertheless, routine and narrow adherence to *p* values and null hypothesis significance testing (NHST) remains staunch in L2 research and its journals (see Plonsky, 2013) despite a number of compelling conceptual and statistical arguments to break this routine.

For the sake of economy, we make three main points to summarize the critiques of NHST and *p* values, in contrast with the relative strengths of effect sizes as an alternative and preferred approach. First, *p* is jointly affected by sample size and the magnitude of the relationship in question and therefore does not reliably reflect the size of its associated effect. Any mean difference between groups or any correlation will reach statistical significance, given a large enough sample. Or as Tukey (1991) put it, "the effects of A and B are always different—in some decimal place—for any A and B. Thus asking 'are the effects different?' is foolish" (p. 100). In other words, the null hypothesis is always a priori false, even though one may not have enough data to reach that conclusion empirically. By contrast, as shown clearly in the arithmetic below, *d* and *r* values are calculated from the available data; these values are not swayed toward statistical significance by a particularly large sample, nor are they deflated by a small one. That said, a *d* or *r* value should be statistically accurate enough—and not just non-null—to reach meaningful conclusions about its practical significance.

Second, *p* values are crude and uninformative in that they encourage dichotomous thinking by classifying results as either significant or not significant. This is fine when the question itself is dichotomous—whether or not to reject the null hypothesis—but researchers should supplement such dichotomous questions with more nuanced judgments about practical significance. Unless one is content with advancing theory and informing practice by asking and answering only yes/no questions about fictitious null populations, a heavy focus on *p* values will be at the expense of overlooking other more useful research results that contribute to a more substantive understanding of how languages are learned. On the latter point, a standardized effect size gives an estimate of the extent to which two variables are actually related (i.e., the magnitude of an effect), which is far more informative when based on reasonably large samples. (A Bayesian approach, wherein observed results are weighed against the predictions of theory and prior findings, could be an additional improvement.)

And third, the cutoff for determining whether or not *p* values are statistically significant is completely arbitrary. Permitting a false significance (Type I) error rate of one in 20 (i.e., alpha $< .05$) is commonly accepted, but there is nothing inherent in this ratio that indicates trustworthiness: "surely, God loves the .06 nearly as much as the .05" (Rosnow & Rosenthal, 1989, p. 1277). By contrast, a *d* value reflects the mean difference between two groups in standard deviation units. Effect sizes such as *d*, *r*, and others (e.g., odds ratios) are standardized metrics that, critically, allow for results to be compared and combined across similar studies via meta-analysis. The formula for the *d* value is:

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

where the numerator is the mean difference and the denominator is either the pooled standard deviation (when the two groups have equivalent variances) or the standard deviation of one of the groups (usually the control or baseline group). Thus, the *d* value expresses the mean difference in terms of the standard deviation across groups or within a reference group across two points in time. When studies make use of the same measure, then means and mean differences are often meaningful in the original raw (unstandardized) metric as well (Bond, Wiitala, & Richard, 2003).

This feature of meta-analysis—the ability to compare and aggregate effect sizes across studies—is one of its primary appeals as a means for synthesizing quantitative research. In addition, meta-analysis allows for nonsignificant findings to contribute useful information to its aggregated results. This is important because meta-analyses may yield practically and statistically significant results, even when their individual constituent studies are not statistically significant. This important point has implications for (a) current findings of publication bias in L2 research journals (i.e., systematically suppressing nonsignificant findings), (b) policies and practices of L2 research, and (c) the conduct of future research (i.e., the need to focus on accuracy, not just significance, and to focus on moderators of practical significance). These three issues will be elaborated upon below.

In light of these and other advantages inherent to the synthetic approach (over traditional, narrative reviews) such as systematicity, comprehensiveness, and transparency (see Norris, 2012; Oswald & Plonsky, 2010), it is not surprising that the use of meta-analysis in L2 research has expanded exponentially in the last decade (Norris & Ortega, 2010), following the footsteps of other research disciplines such as education, psychology, and medicine (Cooper & Hedges, 2009). More precisely, we are aware of only six meta-analyses of L2 research published prior to the year 2000, five from 2000–2004, 28 from 2005–2009, and 58 from 2010–present.

Along with the informational richness of meta-analytic data sets, however, comes the ethical imperative to provide an appropriate and contextualized interpretation of the results (Oswald & Plonsky, 2010), especially given the visibility and high citation rate of meta-analytic findings (Cooper & Hedges, 2009). One of the goals of this article is, then, to inform interpretations of past meta-analytic effects in L2 research and to improve the process of meta-analysis and communication of meta-analytic results in the future. We will take an empirical approach to this goal in Part I, where we report on a synthesis of effect sizes observed in L2 research that leads us to propose a field-specific scale for interpreting effect sizes.

Editorial policy has also been a force that has encouraged the reporting of effect sizes (cf. Matthews et al., 2008). The *Publication Manual of the American Psychological Association* (6th ed.) states that "it is almost always necessary to include some measure of effect size" (APA, 2010, p. 34), and at least four L2 journals have adopted similar policies: *Language Learning* (initially Ellis, 2000, reiterated by DeKeyser & Schoonen, 2007), *Language Learning & Technology*, the *Modern Language Journal*, and *TESOL Quarterly* (see Chapelle & Duff, 2003). Consequently, the practice of reporting effect sizes in L2 research appears to be increasing rapidly (Plonsky, 2014; Plonsky & Gass, 2011), following the recommended practices in psychology (Cumming, 2014).

However, the practice of calculating and reporting effect sizes to date has been relatively mindless and mechanical—a check off the list in author submission guidelines—and L2 researchers typically fail to supplement effect-size reporting with substantive interpretation. Meta-analysis in L2 research has typically had more of a beneficial focus than individual studies on the practical significance of its effect sizes, but the emphasis has mainly been on mean effects and much less so on investigating moderators that capture at least some of the variance (or heterogeneity) across studies. Individual studies should likewise be concerned about the calculation and interpretation of standardized effect sizes as an index of practical significance and whether the effects obtained are robust (or sensitive) to different samples, different measures, different times, and so forth. Thus, beyond distilling a field-specific scale for interpreting effect sizes via our empirical study in Part I, the second equally important goal of our present work is to raise awareness in L2 research about broader conceptual considerations that can usefully assist in interpreting effect sizes and magnitude for both individual studies and for meta-analyses. We will address this second goal in Part II of the article, where we outline eight key considerations for gauging relative magnitude and practical significance of standardized effect sizes.

Part I: Developing an Empirically Based, Field-Specific Scale of Effect Sizes in L2 Research

To move L2 researchers beyond Cohen's (1988) "t-shirt" (Kline, 2009, p. 172) benchmarks and provide more informative interpretations of their results, we now present a survey of effect sizes extracted from 346 primary studies and 91 meta-analyses. This work builds on and extends Oswald and Plonsky's (2010) review, partially replicating meta-syntheses in other social sciences that have sought to better understand the distribution of effects in their respective fields (e.g., Richard, Bond, & Stokes-Zoota, 2003; see also Lipsey & Wilson, 1993). The set of research questions we address are:

1. How large do mean differences (i.e., *d* values) and correlations (*r*) tend to be, as observed in L2 research?
2. How much variance is observed across studies in these reported effect sizes? Specifically, to what extent does the range of observed effects align with Cohen's (1988) standards for small, medium, and large effects??

3. How do *d* values vary by research design (within vs. between subjects; experimental vs. nonexperimental/observational designs) and research setting (lab vs. classroom-based settings)?

Method

The three research questions were addressed by capturing *d* and *r* values in two complementary bodies of quantitative L2 research: 346 primary studies and 91 meta-analyses. As explained in this section, the techniques used to retrieve, code, and analyze these studies are characteristic of research synthesis and meta-analysis. The full list of studies included in the meta-analysis can be found in the Supporting Information online.

Data Collection and Analysis: Primary Effects

Effect sizes from primary study reports were obtained as part of a larger investigation (Plonsky, 2011b) in which 606 quantitative studies published 1990–2010 in *Language Learning* and *Studies in Second Language Acquisition* were collected and coded for effect sizes as well as different designs, analyses, and reporting practices. In total, we obtained 346 samples, where *d* values were reported in (or extracted from) between-groups contrasts in 236 primary reports, and *r* values we obtained from 175 reports. A number of studies contained both types of effect sizes, and the total sample of primary studies contributing one or both types of effect sizes was 346. See Plonsky (2011b, 2013) for additional information on the parameters used for inclusion/exclusion of primary studies.

Once collected, individual study effects were averaged so that each report would contribute a maximum of one *d* and/or *r* to the analysis. Although this practice is typical in meta-analysis (see Lipsey & Wilson, 1993), we remain sensitive to the concern that averaging effects ends up suppressing conceptual and empirical heterogeneity that researchers were originally interested in (for a recent example of modeling heterogeneity in meta-analysis, see Oswald, Mitchell, Blanton, Jaccard, & Tetlock, 2013). However, because the present meta-analysis is a general and descriptive reporting of the L2 literature at large (vs. an inferential analysis for a specific substantive research question), it was critical to the first author for our findings to treat studies equally, including those with a large number of subgroup contrasts (Cameron & Pierce, 1996; Gleser & Olkin, 2009; Hedges & Olkin, 1985; Yirmiya, Erel, Shaked, & Solomonica-Levi, 1998).

Another extremely critical decision point was that we converted negative effects to their absolute value. In an ordinary meta-analysis, this practice is to be avoided; however the central purpose of the current meta-analysis was to determine the magnitude as opposed to direction of effects in the field. In many cases, researchers tested two-tailed hypotheses, and we chose to include the data from these studies rather than exclude them on the basis of having to arbitrarily assign a direction to their effects. Related to this point, many studies with directional hypotheses were proposed in different directions, such as those with both self-paced reading or response time instruments (i.e., tests where the assumption is that a *lower* score indicates greater knowledge of a particular structure) as well as more typical language assessments (i.e., tests where the assumption is that a *higher* score indicates greater knowledge).

Keeping in mind this central purpose of assessing absolute magnitude, descriptive statistics were then calculated for the sample of 236 *d* values and 175 *r* values as a whole and across major designs (experimental vs. nonexperimental/observational, in the case of mean difference effects) and settings (laboratory vs. classroom).

Data Collection and Analysis: Meta-Analytic Effects

Meta-analytic *d* and *r* values were collected from 91 published and unpublished L2 meta-analyses, defined as the aggregation of effect sizes from multiple primary studies of L2 learners. This set of reports, available in Appendix S1 in the online Supporting Information, includes articles, book chapters, conference presentations/posters, dissertations/theses, technical reports, and one unpublished manuscript.

Several strategies were employed to identify this sample, which we believe to be nearly the entire population of L2 meta-analyses conducted to date. We first consulted and collected all studies from two recent reviews of the use of meta-analysis in L2 research, one with an historical perspective (Norris & Ortega, 2010) and the other with a methodological perspective (Oswald & Plonsky, 2010). Additional studies were identified in the syllabi of recent courses on meta-analysis taught by applied linguists (John Norris, Lourdes Ortega, Luke Plonsky) and in a bibliography of meta-analytic L2 research maintained by the first author of this study (http://oak.ucc.nau.edu/ldp3/bib_metaanalysis.html). In addition, seven databases were searched using the keywords *second* or *foreign*, *language*, and *meta-analysis* until the output returned only results that duplicated previously collected studies. The databases were Google, Google Scholar, Linguistics and Language Behavior Abstracts, ProQuest Dissertations and Theses, PsycArticles, PsycInfo, and Education Resources Information Center (ERIC). Further references and reports were obtained through the listserv of the AILA Research Network for Research Synthesis and Meta-Analysis and from professional contacts working in this area. The final sample of 91 meta-analyses comprised data from more than 2,203 primary studies and more than 454,442 individual participants. (Total *k*s and *N*s were not reported in 1 and 22 meta-analyses, respectively.) The sample consists of 49 journal articles (54%), 16 conference presentations and posters (18%), 11 dissertations and theses (12%), 9 book chapters (10%), 4 technical reports (4%), 1 unpublished manuscript, and 1 conference proceeding.

Before arriving at the final sample, several meta-analyses were considered but could not be included for one or more reasons. For example, when multiple versions of the same study were found, or when the same study was available as both a dissertation and a published article, the most recent and/or published version was included (e.g., Grgurović, 2007; Grgurović, Chapelle, & Shelley, 2013; Nakanishi, 2014; Nakanishi, in press; Yun, 2011a, 2011b). A small number of studies were excluded because the aggregated data consisted of reliability estimates rather than validity estimates (e.g., Watanabe & Koyama, 2008) or other indices could not be converted to a *d* or *r* value or that would not apply widely to other subdomains of L2 research. Such indices include variance estimates (e.g., Huang, 2009), likelihood ratios (Dollaghan & Horner, 2013), percentages (Brown, 2014), and differential item functioning units (Koo, Becker, & Kim, 2014). In addition, several reports included the term "meta-analysis" in the title but were qualitative (e.g., Peterson, 2010; Roessingh, 2004; Schwienhorst, 2002) or they included data from only a single study (e.g., Barclay, 1983); these reports were excluded as well. Finally, two meta-analyses were identified as possibly meeting our inclusion criteria, but we were not able to obtain them via library loan or by contacting the authors directly (Diao, 2013; Reljić, 2011).

Once the sample of studies was collected, overall effects for between-group ($k = 67$) and within-group ($k = 25$) contrasts were coded for each meta-analysis. Due to the presence and effect of pre–post correlations in within-group designs but not in between-group designs (Cheung & Chan, 2004; Gleser & Olkin, 2009; Norris, 2012; Plonsky & Oswald, 2012) and due to differences in study designs, we decided to keep the within-group and between-group effects separate (though see Morris & DeShon, 2002). Some meta-analyses reported results for both first-language and L2 studies (Biber, Nekrasova, & Horn, 2011; Goodwin & Ahn, 2010; In'nami & Koizumi, 2009; Jun, Ramirez, & Cumming, 2010). When results were presented separately for L2 studies/learners, only these effects were recorded; otherwise, the study was excluded (e.g., Chiu & Pearson, 1999). Multiple effects were coded in cases when meta-analytic effects from separate groups of studies were reported (e.g., Spada & Tomita, 2010). As

Table 1 Percentiles of mean difference effect sizes (Cohen's *d*) across primary and meta-analytic L2 research

	Percentile				
Type	10	25	50	75	90
Primary[a]	0.28	0.45	0.71	1.08	1.61
Meta-analysis					
Between groups[b]	0.15	0.38	0.62	0.92	1.19
Within groups[c]	0.28	0.61	1.06	1.42	1.69

[a] Based on 236 between-groups contrasts.
[b] Based on 67 meta-analyzed between-groups contrasts.
[c] Based on 25 meta-analyzed within-groups contrasts.

with *d* values obtained from primary studies, and following previous reviews of this type (e.g., Lipsey & Wilson, 1993), these effects were then averaged such that each meta-analysis would contribute a single effect size when calculating summary statistics. In addition, approximately 22% of the sample ($k = 20$) presented meta-analyses of correlation coefficients.

Descriptive statistics were then calculated for the sample of 91 meta-analyses, which yielded a combined total of 67 between-group contrasts, 25 within-group or pre–post contrasts, and 20 meta-analytic correlation coefficients.

Results

Table 1 displays a range of percentile values for each set of *d* values to understand its distributional properties. The between-group contrasts from both primary and meta-analytic studies show a similar pattern. The median *d* value for both is approximately .70, with 25th and 75th percentiles close to .40 and 1.00, respectively. These effects differ from those obtained in meta-analyses of within-group/pre–post designs. For within-subject contrasts, we might expect larger effects, because each participant serves as his/her own control across groups, which reduces error variance and accentuates the strength of the effect (all else being equal). Observed effects resulting from such contrasts are indeed substantially larger, with a median *d* value of 1.06 and 25th and 75th percentiles close to .60 and 1.40, respectively. To complement the numerical display of data in Table 1 and to offer a more complete view of the dispersion of the data, Figure 1 summarizes these results by showing raw data points laid over summary boxplots.

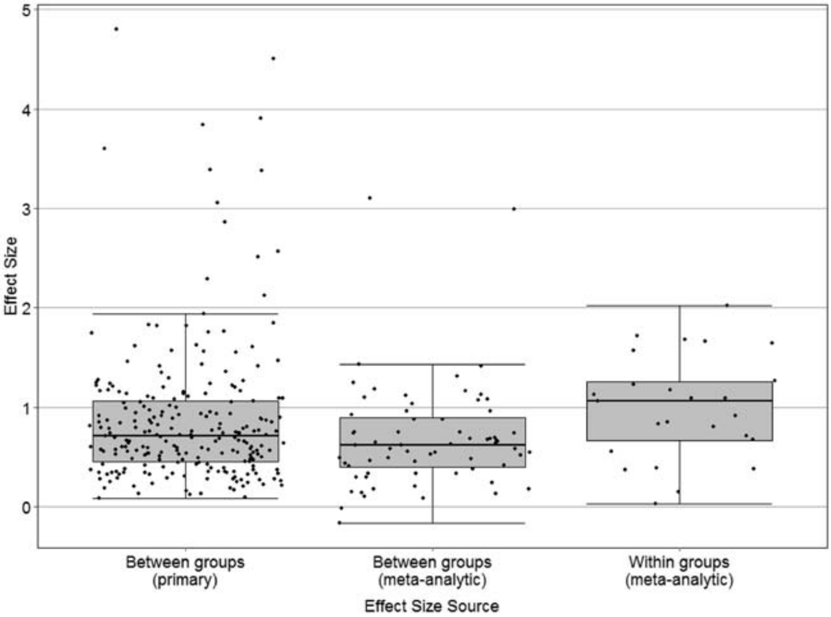

Figure 1 Boxplots of mean difference effects (*d*) from primary and meta-analytic datasets. Data are jittered within category so the number of effects can be seen.

In order to gain a more nuanced perspective on our sample of primary effects, we also combined and averaged primary effects from two major designs (experimental vs. observational/nonexperimental) and settings (laboratory vs. classroom). We found virtually no difference between observational and experimental studies (median $d = .74$ vs. .70, respectively). Likewise and somewhat surprisingly, the data show no relationship between research setting and study outcome at the aggregate level either ($d = .74$ vs. .70 for classroom- and lab-based studies, respectively). Median effects from all four categories are very similar to those of the overall meta-analytic median of the sample ($d = .71$). It is very likely, however, that these results obscure patterns particular to subdomains within the field of L2 research.

In addition to mean differences (*d*), this study was also interested in describing the distribution of correlation coefficients in L2 research. Table 2 summarizes observed *r* values across our sample of 175 primary studies and 20 meta-analyses. Results are strikingly similar. In both samples, the median *r* is .37 with 25th and 75th percentiles close to .25 and .60. And again, in order to provide a fuller description of the range of correlational findings

Table 2 Percentiles of correlation coefficient effect sizes (*r*) across primary and meta-analytic L2 research

	Percentile				
Type	10	25	50	75	90
Primary[a]	.17	.25	.37	.54	.71
Meta-analysis[b]	.18	.25	.37	.68	.83

[a] Based on correlations from 175 primary studies.

[b] Based on 20 meta-analyzed sets of correlations.

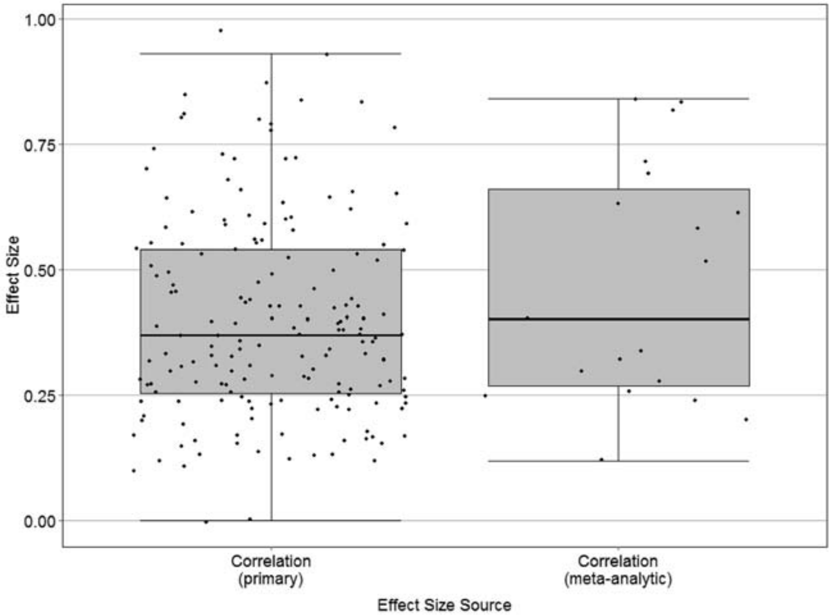

Figure 2 Boxplots of correlation effects (*r*) from primary and meta-analytic datasets. Data are jittered within category so the number of effects can be seen.

in L2 research, Figure 2 presents each primary and meta-analytic correlation laid over summary boxplots.

As with mean difference effects, we also examined the role of setting as a potential moderator of correlational findings. And again we found only a minor difference between *r* values resulting from classroom- and lab-based studies (median = .33 vs. .37, respectively).

Discussion

As described in our review of the literature, quantitative methodologists from a wide variety of fields have argued for decades in favor of more localized interpretations of effect sizes. In this sense, Cohen's (1988) unqualified levels of small ($d = .2$, $r = .1$), medium (.5, .3), and large (.8, .5) have certainly been overused, if not misused, in L2 research as across many other fields in the social sciences. Moreover, the results of the present study provide empirical evidence that Cohen's scale underestimates the range of effects typically obtained in L2 research. We propose the following field-specific and empirically based scale for general descriptions and interpretations of *d* and *r* values:

For mean differences between groups, *d* values in the neighborhood of .40 should be considered small, .70 medium, and 1.00 large. These estimates of (roughly) small, medium, and large effects were chosen based on their approximate correspondence to the 25th, 50th, and 75th percentiles, respectively, for between-group contrasts in primary and meta-analytic research. Our results clearly indicate that Cohen's (1988) labels for small ($d = .2$), medium ($d = .5$), and large ($d = .8$) mean difference should not generally be applied to L2 research; we therefore urge L2 researchers to adopt the new field-specific benchmarks of small ($d = .40$), medium ($d = .70$), and large ($d = 1.00$) in order to interpret the practical significance of L2 research effects more precisely. For *d* values resulting from pre–post or within-group contrasts, effects are generally larger. This difference is due to intragroup correlations found in this type of design. In order to take this into account, we propose an alternate scale that is again based on the 25th, 50th, and 75th percentiles of observed effects. This scale considers a *d* value of .60 as generally small, 1.00 as medium, and 1.40 as large. The difference between observed effects here and Cohen's scale (which was originally proposed for interpretation of between-group contrasts only) is even more pronounced. Based on the effects we observed from within- versus between-group contrasts, we would also urge future meta-analysts to analyze data from these two major designs/contrasts separately.

For correlation coefficients, we suggest that *r*s close to .25 be considered small, .40 medium, and .60 large. These values correspond roughly to the 25th, 50th, and 75th percentiles in our primary and meta-analytic samples. As with mean difference effect sizes, these results show very clearly that Cohen's benchmarks for small, medium, and large correlations (.1, .3, .5) underestimate and are not appropriate for interpreting those found in L2 research.

One major benefit of the summary of effects reported here is the ability of future studies and meta-analyses in particular to look inward and gauge the relative magnitude in individual subdomains of L2 research, rather than

Table 3 Comparison of meta-syntheses in different social sciences

Domain	Average Effect Size (*SD*)
Mean Differences Between Groups (*d*)	
L2 Research	0.69[a] (0.55)
Education[b]	0.40 (0.13)
Psychological, Educational, and Behavioral Treatments[c]	0.47 (0.29)
Educational Technology[d]	0.35 (0.21)
Correlation Coefficients (*r*)	
L2 Research	0.46[a] (0.24)
Social Psychology[e]	0.21 (0.15)

[a] In order to match the analyses of the other fields included here, this value was calculated as the mean, rather than the median, of observed meta-analytic effects.
[b] Hattie (1992), $k = 134$.
[c] Lipsey & Wilson (1993), $k = 302$.
[d] Tamim, Bernard, Borokhovski, Abrami, and Schmid (2011), $k = 25$.
[e] Richard et al. (2003), $k = 322$.

using omnibus benchmarks that may not be as relevant. Looking outward, however, the findings of this study can also be roughly compared to similar reviews in other disciplines to give us a sense of where our field's effects tend to fit within the larger social sciences. From this perspective, we see that quantitative L2 research produces substantially larger effects than those typical in many other fields. Meta-analyses and other syntheses like the present study have been reported in several domains with substantive and methodological ties to applied linguistics, such as education and psychology. Table 3 presents overall results from four such studies along with those of the present study. Average mean difference effects from all three meta-syntheses fall close to .40, approximately one-third of a standard deviation below effects typically found in L2 research. Hattie (1992), for example, synthesized results across 134 meta-analyses in education. His results showed that educational interventions with academic outcomes produce an average *d* value of .40. Similarly, Lipsey and Wilson (1993) observed a median *d* of 0.47 across 302 meta-analyses of psychological, educational, and behavioral treatments. Like the present study, their results also showed an inflation of effects in pre–post ($d = .76$) compared to between-group designs ($d = .47$). As with mean difference effects, Richard et al. (2003) found average meta-analytic correlations from social psychology to be smaller than half of those in the present study.

One explanation for the difference in observed effects across fields might be the role of novelty of L2 research. As a relatively young field (see, e.g., Gass, Fleck, Leder, & Svetics, 1998), the relationships and variables L2 researchers examine may be somewhat coarse compared to education and psychology, leading to less refined contrasts/analyses and therefore generally larger effects. (See discussion below on how this pattern might play out in individual subdomains of L2 research.)

The magnitude of effect sizes observed in L2 research compared with other disciplines might also be related to the combined effects of small samples on one hand, which are typical in our field (Plonsky, 2013; Plonsky & Gass, 2011),[1] and a fixation with statistical significance on the other. Mathematically speaking, in order to offset a small sample's lack of statistical power needed to obtain a *p* value below .05, an effect size must be relatively large. And given the well-documented bias toward studies with statistically significant findings (Rothstein, Sutton, & Borenstein, 2005; Sutton, 2009), it is possible that only those studies with larger effects will be published, leading to inflated overall effects at the primary and meta-analytic levels. Considering these points and their particular relevance in the context of L2 research, we might want to assume the results presented above and those in individual meta-analyses to overestimate true population effects.

We feel the need to emphasize an important point of caution. The benchmarks we have offered as a result of our study in Part I of this article are meant to serve as very general indicators of the magnitude of mean differences and correlations typically observed in L2 research. We do not suggest they be applied indiscriminately as we have seen with *p* values or even with Cohen's (1988) benchmarks. As Thompson (2001) warned, "if people interpreted effect sizes [using fixed benchmarks] with the same rigidity that .05 has been used in statistical testing, we would merely be being stupid in another metric" (pp. 82–83). Rather than view these values as fixed, we recommend that L2 researchers consider numerous additional factors when interpreting their effects, whether they result from primary or secondary analyses. It is for this reason that a host of broader conceptual issues must also be kept in mind when interpreting magnitude in individual studies vis-à-vis the field-wide empirical expectations we have distilled and proposed. This is the goal of Part II of this article.

Part II: Additional Considerations for Interpreting Effect Sizes in L2 Research

Part I focused on a survey of effect sizes observed in L2 research, the distribution of which can be used as a starting point when interpreting quantitative results. In order to make the most of L2 research and to inform L2 theory, practice, and future research most effectively, in Part II of this article we turn to a broader discussion of eight additional factors that ought to be considered when interpreting effect sizes. We hope that a better understanding of these factors will lead to more nuanced and informative reports, which can then more accurately guide the field. We also hope to seize upon the relatively recent introduction and presence of effect sizes in L2 research to encourage and establish more informed research practice in the field. Although other disciplines such as education and psychology have a longer tradition of reporting effect sizes, they also suffer from somewhat rigid conventions with respect to interpretations of these effects, defaulting frequently to Cohen's (1988) benchmarks (Fritz, Morris, & Richler, 2011; Hill, Bloom, Black, & Lipsey, 2008). We see the lack of such established conventions in L2 research as an opportunity; we can learn from and move beyond the ingrained habits of other fields. Although the movement toward practical significance in L2 research is still young, our work can help establish a pattern of more informative interpretations of effects in L2 research.

In the eight subsections that follow, we go beyond the benchmarks described above to provide guidance on interpreting effect sizes by considering: (1) previous studies addressing similar relationships, (2) "literal" (i.e., mathematical) interpretations of *d* and *r* values, (3) the theoretical and/or methodological maturity of the domain in question, (4) research designs and settings, (5) experimental manipulation vis-à-vis practical significance, (6) the presence of publication bias or biases, (7) psychometric artifacts, and (8) methodological quality.

Previous Studies of Similar Relationships

We begin this discussion with perhaps the most important consideration when interpreting quantitative results: Researchers should examine effect sizes relative to the effects of previous studies interested in similar relationships (even without the benefit of information from a meta-analysis). For primary studies, a meta-analysis of the domain or subdomain to which the study belongs, if available, is a great place to start. Recent and comparable primary studies can also give an indication of the relative magnitude of observed effects, and differences and similarities both substantive and methodological should be described and explained (Stukas & Cumming, in press).

Meta-analysts can look to the results of other meta-analyses when explaining their findings. The proliferation of meta-analytic replications in L2 research is particularly helpful here (see Plonsky, 2012b; Plonsky & Brown, in press).

For example, in interpreting the effects of her meta-analysis of the effects of extensive reading, Wang (2010) compared her findings to those of Krashen's (2007) meta-analysis of the same topic. When a meta-analytic replication is not available, reviews of comparable domains can still be put to good use. In the absence of previous meta-analyses of the effects of pronunciation instruction, Lee, Jang, and Plonsky (in press) examined their overall result in relation to other domains of L2 instruction that had been meta-analyzed, including morphosyntax (Goo, Granena, Yilmaz, & Novella, in press; Norris & Ortega, 2000; Shintani, Li, & Ellis, 2013; Spada & Tomita, 2010), vocabulary (Chiu, 2013; Wa-Mbaleka, 2008; Won, 2008), and pragmatics (Jeon & Kaya, 2006). Meta-analytic results from other more established disciplines such as (first-language) education, cognitive science, and individual differences provide still another point of comparison. Plonsky (2011a), for instance, observed that the overall effectiveness of L2 strategy instruction ($d = .49$) aligned closely with Hattie, Biggs, and Purdie's (1996) meta-analytic findings for strategy instruction in first-language educational settings ($d = .45$).

Literal (i.e., Mathematical) Interpretations of Effect Sizes

One relatively straightforward approach to interpreting effect sizes involves considering their literal or mathematical meanings. Mean difference effect sizes such as Cohen's *d*, like *z* scores, are reported in *SD* units and can be interpreted as such. (Recall from above that the *SD* of one or both groups is the denominator in the formula for Cohen's *d*.) A *d* of .50 therefore indicates that one group scored on average one-half of a standard deviation above the other group. (For examples of such interpretations in L2 research, see Won, [2008] and Taylor, Stevens, and Asher [2006], among others.)

Because *d* values are essentially *SD* units, we can also interpret them as the percentage of control/comparison group participants in the given study that fell below the average experimental group participant (assuming a normal distribution in both samples). A *d* of 0 then indicates that the mean of the experimental group is equal to the 50th percentile (or median) of the control group; a *d* of .40, shown above to be a somewhat small effect in L2 research, indicates that the mean of the experimental group is at the 66th percentile of the control group; a larger effect of 1.00 indicates that the mean of the experimental group is equal to the 84th percentile of the control group, and so on. In'nami and Koizumi (2009) took this approach among others when interpreting the results of their meta-analysis of test format effects. The authors inferred that multiple-choice test takers would be 78% more likely to score above the average on an open-ended test of L2 listening. For further discussion on converting *d* values to fairly easy-to-interpret percentile-based units, see Cohen's (1988) U_3, and Rosenthal, Rosnow, and Rubin's (2000) Binomial Effect Size Display.

Similarly, η^2 and R^2 effect sizes can also be interpreted in relatively straightforward terms. These indices tell us the percentage of shared variance between two continuous variables. For instance, the median correlation from the present primary and meta-analytic sample, $r = .37$, could be squared and interpreted as showing that, on average, correlated variables in L2 research share 14% of their variance. In the case of categorical predictor variables, these effect sizes indicate the percentage of variance in the dependent variable that can be accounted for by group membership. In other words, they tell us how well an independent variable explains differences among groups.

Domain Maturity and Changes in Effects Over Time

There is no reason to assume that effects in any particular domain would remain static over time.[2] In fact, we might expect dynamics to occur, as is found in the Proteus effect, where large and interesting effects are reported early in the evolution of an area of research, but then these effects regress toward the mean effect—and possibly toward nonsignificance—as the phenomenon is investigated more carefully (for observation of the Proteus effect in the medical field, see Trikalinos et al., 2004; Trikalinos & Ioannidis, 2005). In addition to the Proteus effect, identifying any time-based patterns of effects can help meta-analysts to more thoroughly illustrate findings in their domains of inquiry.

A first scenario in which effect sizes might change over time involves a trajectory of research outcomes for a particular domain in which methodological adjustments or improvements lead to larger effect sizes over time (Fern & Monroe, 1996). In experimental L2 research, such changes may result from the realization that longer or stronger treatments produce larger differences between groups. For instance, though perhaps counterintuitive, Yang (2014) found shorter study abroad programs to yield larger effects on average than longer ones. An increase in effect sizes might also be found when the psychometric properties of instruments, the standards for which are generally lower in an emerging research area (Brutus, Gill, & Duniewicz, 2010), are refined over time. A shift in the type of measures used can also contribute to changes over time. For example, Li (2010) and Lyster and Saito (2010) both reported larger effects for open-ended tests, a format found to be increasing over time in both Mackey and Goo (2007) and Spada and Tomita (2010).

In contrast with the scenario just described, an alternate and somewhat opposite pattern of effects may also play out in a body of empirical literature characterized by theoretical developments and maturation taking place over time. Early research in a given area is often characterized by strong manipulations that set out to determine whether an effect exists and thereby determine whether the claims of a particular and usually novel hypothesis merit further attention. Such experiments would tend to yield larger effect sizes (Ioannidis, 2008; Kline, 2004). Subsequently, after an effect is found, research efforts may shift to the generalizability of an effect across samples, settings, tasks, and so forth (see Plonsky, 2012b; Stukas & Cumming, in press). More mature domains are therefore more likely to be examining relationships qualified by a more specific situation or criterion. In domains where this scenario is observed, theoretical maturity could be inversely correlated with outcomes, and a decrease in effect sizes would be obtained over time. In their review of the interactionist tradition in L2 research, Plonsky and Gass (2011), for example, observed a trend toward smaller effects and a narrowing of confidence intervals: *d* values across the 1980s, 1990s, and 2000s averaged 1.62 [Confidence Intervals (CI): 0.99–2.25], 0.82 [CI: 0.60–1.04], and 0.52 [CI: 0.36–0.69], respectively. They interpreted this finding as evidence of a move toward more refined relationships and analyses and thus of the domain's increasing maturity as well.

Of course it is also quite possible for both of these scenarios to play out simultaneously. In these cases, simply calculating mean effects across decades or even correlations between year of publication and effect size may obscure the actual trends taking place. Estimates of variance/error (e.g., CIs) for effect sizes must be examined as well, and visual displays of effect sizes involving moderator categories can be especially useful (Schild & Voracek, 2012; Stukas & Cumming, in press).

Combs (2010) provided an alternate perspective on the latter scenario (i.e., a decrease in effects over time). In his review of effect sizes (correlation coefficients, in this case) for organizational research reported in the *Academy of Management Journal*, he found the magnitude of effects to be decreasing over time. However, instead of attributing that change to empirical demonstrations of theoretical nuance, he argued that the inverse relationship between date of publication and effect size was related to the increase in average sample size that took place during the period in question. More precisely, he claimed that as reviewers and editors began to recognize the importance of statistical power and required larger samples, contributing authors were able to obtain and publish $p < .05$ for smaller correlations. Statistically speaking, from the perspective of null hypothesis testing, it is hard to counter this argument. (Holding a correlation constant, regardless of how small it is, $p < .05$ can always be attained given a large enough sample.) Combs' approach, however, deemphasizes two important factors. First, we should expect that more mature domains might ask more varied questions that might produce more heterogeneous effect sizes overall, whether or not their sample sizes have increased. And second, related to this first point, meta-research at the field-wide level, such as the results of Part I of this article, is often blind to the variance across multiple subdomains in the sample. We would urge Combs and others to consider both statistical and substantive developments that account for changes in effects. Nevertheless, his study presents a worthwhile example of an exploration of effect sizes over time from another discipline.

Unlike meta-analyses examining findings within a single subdomain of L2 research, the field-wide scope of the current study's analysis of primary effects, like Combs (2010), may blur a meta-analyst's ability to detect whether one, both, or neither of these patterns has occurred. In other words, there is no doubt that the last few decades of L2 research have seen numerous domains rise, fall, mature, improve, decline, diverge, and so forth, but these patterns may not leave observable statistical traces in the aggregate. The individual domains that have been meta-analyzed and that have provided effect sizes for the primary studies in their samples, however, present an interesting opportunity to reanalyze data and examine the extent to which either or both of the scenarios described here have occurred.

Finally, in domains where one of these patterns is observed, meta-analysts should also examine whether changes in reporting practices have also occurred. The combined presence of these two changes may constitute a source of bias in overall effects. Since the introduction of meta-analysis to L2 research, the field has seen an increase in reporting effect sizes or the descriptive statistics required to calculate effect sizes (Mackey & Goo, 2007; Plonsky, 2014; Plonsky & Gass, 2011). And as we mentioned earlier, four L2 journals—*Language Learning*, *Language Learning & Technology*, the *Modern Language Journal*, and *TESOL Quarterly*—have aligned themselves with the APA in requiring the reporting of effect sizes. Although these are no doubt steps in the right direction, the shift toward greater transparency in primary studies and, thus, a disproportionately recent set of studies that contribute effect sizes, may introduce both upward and downward forces of bias at the meta-analytic level. It is, of course, important to understand the nature and magnitude of such biases rather than accepting on faith that biases do not matter or will cancel out in the empirical wash.

Research Setting and Design

Considering the role of research setting and design may also help explain variability in primary and meta-analytic effects. At least two contextual contrasts are worthy of attention: (a) laboratory versus classroom and (b) second versus foreign language setting.

Depending on the variables of interest, studies carried out in laboratories and classrooms might produce different effects. In the lab, researchers can often exert greater rigor and experimental control over environmental and other potentially contaminating variables, enabling greater isolation of—and thus larger—intended effects (see Cohen, 1988, p. 25). This pattern has been observed in meta-analyses of several domains of L2 research. In Li's (2010) meta-analysis of corrective feedback, for example, the average effect for lab-based studies ($d = 1.08$) was more than twice that of classroom-based studies ($d = .50$; see similar results in Mackey & Goo, 2007; Lee et al., in press; Plonsky, 2011a; for a discussion and a rare and important example of a primary study that compares treatments across lab and classroom settings, see Gass, Mackey, & Ross-Feldman, 2005).

Some domains may also shift over time toward classroom-based studies as researchers seek to establish ecological validity for findings previously established in the relatively controlled vacuum of the lab (e.g., in the domain of L2 oral interaction; see Plonsky & Gass, 2011). We describe this pattern to alert synthesists to the possibility of an interaction between simultaneous shifts in context and in the size of the effects obtained (see also previous section).

Another contextual variable, second versus foreign language setting, may also moderate effects. As in lab versus classroom comparisons, the sensitivity of certain domains to research setting may vary both in direction and magnitude. Whereas Cobb's (2010) meta-analysis of task-based interaction found a strong advantage for studies carried out in foreign-language settings ($d = 0.89$ vs. 0.14 in L2 settings; see similar results in Li, 2010), the opposite was observed in Taylor et al.'s (2006) meta-analysis of reading strategy instruction ($d = .44$ in foreign- language vs. .63 in L2 settings; see also Plonsky, 2011a).

When making sense of results such as these, as throughout the synthetic process, we cannot overemphasize the value of the substantive reviewer's expertise and familiarity with the domain in question. For example, Lee et al.'s (in press) meta-analysis of pronunciation instruction found effects based on controlled outcome measures (e.g., reading a word list) to be substantially larger than free or open-ended ones (e.g., picture description task). They explained this difference to likely be the result of the lack of communicative value in controlled tasks, which allows participants to focus more on their pronunciation. Free tasks, by contrast, require participants to focus simultaneously on what they say and how they say it.

A number of additional design features can also be considered with respect to their influence on study outcomes. The study reported in Part I presents compelling evidence of one such feature: within- vs. between-group contrasts as the basis for estimating the *d* value for a particular study or set of studies (see Norris, 2012). Moreover, inflated effects were not only observed in the aggregate. A number of individual meta-analyses in our sample preserved this distinction in their presentation of summary results, almost all of which (18 out of 19) found pre–post contrasts to produce larger effects than those derived from control–experimental contrasts. This makes sense in that for within-group (repeated measures) designs, participants serve as their own control, thus accounting for more error variance (i.e., a main effect for subjects in an analysis of variance model).

In the case of between-group designs where a control group is used in the calculation of effect sizes, primary and secondary researchers should further consider the *type* of control condition. Some experimental domains or traditions may regularly include true control groups that receive no intervention at all, leading to starker contrasts and larger effects. Due to practical priorities, or due to the constraints of logistics or ethics, others may only include comparison groups (i.e., those that receive an alternate or traditional treatment). At the secondary level, the meta-analyst can examine these differences to help isolate true experimental effects resulting from the treatment in question (Norris, 2012).

Experimental Manipulation Vis-à-vis Practical Significance

The meta-analysis movement currently taking place in applied linguistics has raised researchers' awareness of the importance of practical significance (see, e.g., Loewen et al., 2014; Norris & Ortega, 2006). In contrast with statistical significance (i.e., *p* value), practical significance is usually expressed as an effect size (i.e., magnitude) and provides researchers and practitioners with information to make judgments about the substantive importance of a particular research finding or relationship.

The move toward habitually reporting and discussing practical significance is clearly, in our opinion, a move in the right direction. Often overlooked in such discussions, however, are the efforts required to induce experimental manipulations that are deemed to be practically significant. In other words, the potential benefits of a given treatment must be weighed against the costs: time, resources, effort, experimental manipulation, and so forth (see Rosenthal & Rubin, 1979; Stukas & Cumming, in press). And the benefits of an experiment may not simply be reflected in the change of one variable; an experiment might have effects on student performance, but also student motivation and persistence, teacher self-efficacy, and so on.

In some cases, the practical significance of the research outcomes clearly justify the intervention. Lee and Huang (2008) and Alsadhan (2011) separately meta-analyzed the effects of input enhancement on grammar learning and noticing. Both studies showed that such treatments arrive at effects that would by Cohen's (1988) standards be considered small: $d = .22$ and .30, respectively. However, full consideration of the practical significance of these results takes the critical reader beyond the numerical outcomes to an assessment of the means required to achieve them as well. The minimal manipulation and effort required to induce these effects (e.g., bolding or highlighting grammatical features in source input), coupled with their potential to aid in L2 development, would most likely support implementation of this technique at some level because the investment of time, effort, and money is low, and the effects may be persistent and beneficial across L2 learning environments.

In other cases a cost–benefit analysis may call into question the use of time and other valuable resources. For example, Wang (2010) meta-analyzed the effects of extensive reading on several outcome/dependent variables (e.g., reading comprehension, reading speed, writing fluency). The overall effectiveness across all outcome types for pre–post contrasts was estimated at a *d* of 1.13, which corresponds roughly to a medium effect in the realm of L2 research, according to the criteria we have proposed. If one takes into account the duration of the treatments—often spanning one or more entire semesters and taking up dozens of hours of valuable class time—however, practitioners and policy makers would likely temper their enthusiasm for these results.

With these issues in mind, a small number of meta-analyses of L2 research have explored treatment length as a moderator of study outcomes. As we might expect, Jeon and Kaya (2006, pragmatics instruction), Lyster and Saito (2010, classroom feedback), Plonsky (2011a, strategy instruction) all found that longer interventions led to larger effects (cf. Norris & Ortega, 2000; Yang, 2014). On the one hand, positive correlations between length of instruction and study outcomes could be used to argue for longer treatments or for the inclusion of longer pedagogical units in L2 curricula. It can also be argued, however, that small positive correlations between treatment lengths and outcomes indicate potential for greater efficiency in terms of curricular design and policy. Finally, in order to put a finer point on analyses such as these, future meta-analyses should consider not only length in days or weeks but also intensity (i.e., hours or target items per week; Norris, 2012).

Publication Bias

In the most general sense, publication bias is present when the sample of primary studies to be meta-analyzed is not representative of the entire population of interest (this bias is similar to concerns about unrepresentative sampling within individual studies). There are several sources of publication bias in the social sciences, but the most well documented and well known occurs as the result of the preference among authors, reviewers, and editors to publish only those studies with statistically significant results (Norris & Ortega, 2000; Rothstein et al., 2005; Sutton, 2009). From the meta-analytic perspective, this particular type of publication bias yields an inflated overall effect, because larger effects are more likely than smaller effects to be statistically significant.

Despite a lack of systematic investigation in this area, there is growing evidence of publication bias among L2 meta-analyses that have investigated this issue. In one of few such studies, Lee and Huang (2008) grouped and compared the effects of textual enhancement among (a) published results (not based on a dissertation; $d = .55$, $k = 8$), (b) published results based on a dissertation ($d = .24$, $k = 4$), and (c) unpublished dissertation results ($d = -.01$, $k = 4$). Not only were published findings larger than dissertations, but dissertation results that were published were also substantially larger than those in unpublished dissertations. Plonsky (2013) found further evidence of publication bias in L2 research in numerous omissions of results ($k = 27$) in cases where $p > .05$. Furthermore, due to the small and underpowered samples typical of L2 research, small effects in the population may be particularly prone to overestimation, as in the textual enhancement example above (i.e., a Type M error, or magnitude error; see Gelman & Weakliem, 2009; Plonsky, 2013, 2014; Plonsky, Egbert, & LaFlair, in press; Plonsky & Gass, 2011).

L2 meta-analysts can combat the threat to the validity of their results posed by a biased sample of effect sizes, first, by thoroughly examining the presence of publication bias as part of the standard process of exploring the meta-analytic dataset. This can take the form of a moderator analysis based on publication type/status (as in the example by Lee & Huang, 2008, just described), the use of graphic displays of data (e.g., funnel and forest plots, as seen in Norris & Ortega, 2000), or, ideally, a combination of statistical and visual analyses (as done by Li, 2010). (For a review of options for displaying data visually, see Anzures-Cabrera and Higgins [2010].)

A second and possibly more effective means toward reducing bias is inclusivity in the search for primary studies. Whenever possible, unpublished research including dissertations, conference papers/posters, ERIC documents, working papers, and so forth should be included. Some meta-analysts have opted to only include published or so-called high-quality research, citing what is known in the meta-analysis literature as the garbage in, garbage out (GIGO) validity threat (e.g., Truscott, 2007). We would argue, like Plonsky and Brown (in press), that a more inclusive approach provides a fuller picture of substantive results. Furthermore, the results of more comprehensive syntheses also enable the reviewer to examine methodological quality across the domain, providing (a) the opportunity to analyze study features in relation to outcomes (see Section 8) and (b) an empirical basis for recommendations to be made for future research. In a meta-analysis of technology-enhanced language learning, Zhao (2003) limited his search to five source journals that frequently publish research on computer-assisted language learning (CALL). Not only did this choice provide a narrow view of the domain (as there are certainly other journals and nonjournal venues where CALL research is found), but it may also introduce the potential for a substantive bias. That is, the main journals in this area, given their invested interest in CALL, may be more likely to publish research that favors findings supporting the role of technology in language learning.

Other differences between journals may also lead to a biased sample of primary studies. When differences in effects among journals coincide with stricter or more lenient reporting practices (e.g., effect sizes or descriptive statistics needed to calculate effect sizes), the sample of primary studies will be biased toward the results available in the journal with more thorough reporting (for a comparison of methodological features, reporting practices, and effects in *Language Learning* and *Studies in Second Language Acquisition*, see Plonsky, 2014). Our hope is that in the era of open access and online journals, all journals will be leaning toward more transparent and thorough sharing of data and results, thus minimizing this type of bias in the future.

Instrument Reliability and Other Psychometric Issues

The properties of different measures used in L2 research are often overlooked both conceptually, in terms of how theoretical constructs might be operationalized differently, as well as psychometrically, in terms of their statistically biasing effects on reported effect sizes. Conceptually, measures might vary on many factors that meta-analysis is able to examine, so long as there are enough effect sizes available. These measurement factors might include (but are not limited to) the language in task prompts or instructions, who provides the data on the measure (e.g., student, teacher, peer), and the medium (e.g., Web based vs. paper and pencil vs. oral). Psychometrically, even the most basic meta-analyses will take the sample size N into consideration, where studies with larger Ns contribute more to meta-analytic results than studies with smaller Ns.

However, this statistical weighting will not identify the need to subdivide an overall meta-analysis into meaningful subgroups, such as those defined by the aforementioned measurement factors.

Certain methods of meta-analysis can also incorporate psychometric information into the analysis and interpretation of results in an attempt to correct for error and bias in observed effects relative to the population effects of interest (Hunter & Schmidt, 2004). For example, items that measure the same construct will tend to have high alpha reliability (although the opposite may not be true: high alpha reliability does not guarantee that items measure the same construct). A psychometric approach to meta-analysis will weight effect sizes such that effects measured more reliably (i.e., with higher alpha coefficients) receive more weight and therefore contribute more to the meta-analytic average, and lower alphas contribute less to the meta-analytic average. Even though psychometric corrections in meta-analysis will tend to increase the magnitude of observed effects more when reliability is lower, effect sizes with lower reliability are also weighted less in the meta-analysis, all other factors equal.

Range restriction is another factor often overlooked in both primary studies and meta-analyses. For example, the *d* values in an intervention with atypical samples such as gifted-and-talented students, true beginners, or near-native speakers, might be smaller (range-restricted) compared with the larger population of learners to which one seeks to generalize, because these samples are likely to have a very restricted range of cognitive ability and/or performance. Psychometric meta-analysis can correct the effects of primary studies by how much range restriction occurs. Knowing the range of cognitive ability in the restricted sample and in the full population allows one to correct the magnitude of the *d* value to be estimated in the full population. What is also required is the correlation between cognitive ability and the outcome on which the *d* value is based. Range restriction correction formulas are complex, because there are various mechanisms for range restriction, various types of data available, and therefore various possible corrections that can be applied (Sackett & Yang, 2000). But arguably, endeavoring to make these corrections in L2 settings where range restriction is present might better equalize the studies in the meta-analysis and lead to better substantive understanding of effects of interest at both the primary and meta-analytic levels. Otherwise, variation in the observed effects might not be largely for theoretically expected reasons but instead due to variation in the reliability of the measures and the selectivity of the sample. (For examples of L2 meta-analyses that have made corrections for range restriction, see Jeon & Yamashita, 2014; Masgoret & Gardner, 2003; Ross, 1998.)

Methodological Quality

A final factor to consider when interpreting effect sizes is the nature and possible relationship between research practices and study effect sizes. This type of analysis begins with the assumption that "study results are determined conjointly by the nature of the substantive phenomenon under investigation and the nature of the methods used to study it" (Lipsey, 2009, p. 150). Put another way, "effect sizes are not magically independent of the designs that created them" (Vacha-Haase & Thompson, 2004, p. 478).

Several meta-analyses have investigated the methodological features across studies, particularly those associated with quality and data reporting practices, in relation to outcomes. Russell and Spada (2006), for instance, calculated the average effect of error correction based on whether studies reported the reliability and validity of their dependent measures. Although their results showed larger effects for studies that did not report estimates of reliability or validity, the opposite pattern has been observed in several subsequent reviews (e.g., Adesope, Lavin, Thompson, & Ungerleider, 2010; Plonsky, 2011a). Other more specific features of methodological quality analyzed in relation to effect sizes at the meta-analytic level include pretesting and delayed posttesting, random assignment to experimental conditions, control for bias, and whether or not different types of data were reported in the primary studies, such as effect sizes, confidence intervals, and exact *p* values (Adesope et al., 2010; Adesope, Lavin, Thompson, & Ungerleider, 2011; Grgurović et al., 2013; Jun et al., 2010; Plonsky, 2011a, 2011b; Plonsky & Gass, 2011).

The findings from these studies provide partial support for a relationship between different research practices and study outcomes (see Lipsey & Wilson, 1993; Prentice & Miller, 1992; Wilson & Lipsey, 2001). However, no one would claim, for instance, that it is the act of randomly assigning participants to experimental conditions or reporting reliability that causes larger or smaller effects per se (see Lipsey, 2009). Other concomitant causes are usually likely. For example, it is entirely possible that studies with random assignment to groups are more likely to be carried out in lab contexts where researchers exercise greater experimental control and are thus able to obtain larger effects. Similarly, studies using highly reliable instruments might be more likely to both report reliability and to obtain larger effects because their results will not be attenuated by low reliability. Researcher experience may be a critical third variable here that drives the relationship between researcher knowledge of which variables to study (substantive expertise) and how to study them (methodological expertise) which, in tandem, may lead to the typically observed relationship between (higher) study quality and (larger) effect sizes.

More generally, we argue that exploring relationships between study features and outcomes not only quantitatively but also *qualitatively* can help the research community to understand more clearly how L2 research is carried out. In other words, even when the myriad features associated with study quality in a given domain do not leave a statistical trace through its quantitative relationship with the size of effects, results pertaining to methodological rigor still contribute to our interpretation of the quality of the research giving rise to those effects. For this reason, primary researchers should examine, report, and evaluate study characteristics in the context of their respective domains, and meta-analysts should do so as well. For guidance on and existing instruments for defining and operationalizing quality in primary research, we recommend consulting the APA Publications and Communications Board Working Group on Journal Article Reporting Standards (2008), the Design and Implementation Assessment Device (DIAD; Valentine & Cooper, 2008), and in the context of L2 research, the recommendations found in Larson-Hall and Plonsky (in press) and Plonsky (2014).

Conclusion

Effect sizes have much to offer in terms of more precisely informing theory and practice in L2 research. Their potential, however, can only be more fully realized by moving beyond one-size-fits-all interpretive rules of thumb and instead understanding and applying discipline-specific benchmarks for interpreting results from primary and secondary research. Toward this end, we have collected and summarized the distribution of mean difference and correlational effects observed in 346 primary studies and 91 meta-analyses, the distribution of which we have used to propose discipline-specific benchmarks for interpreting effect sizes in L2 research. We have also outlined a set of additional key issues worthy of attention when interpreting L2 effects. We hope that, when taken together, these guidelines and considerations will contribute to meaningful and efficient advances in the field.

Final revised version accepted 6 April 2014

Notes

[1] Simply put, the populations of interest in educational and psychological research are often, by definition, larger than those of L2 research. Our field, it should be noted, has also been slow to recognize the importance of statistical power (see Crookes, 1991; Plonsky, 2014).

2 This same point could also be applied to the larger field of L2 research and to the benchmarks proposed here. Although there is no evidence yet for a field-wide trend toward smaller or larger effects, as effects in the field continue to accumulate, we will need to revisit and refine these benchmarks.

References

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.

Alsadhan, R. O. (2011). *Effect of textual enhancement and explicit rule presentation on the noticing and acquisition of L2 grammatical structures: A meta-analysis*. Unpublished doctoral dissertation, Colorado State University, Fort Collins, CO.

American Psychological Association. (2010). *Publication manual of the American Psychological Association* (6th ed.). Washington, DC: Author.

Anzures-Cabrera, J., & Higgins, J. P. T. (2010). Graphical displays for meta-analysis: An overview with suggestions for practice. *Research Synthesis Methods*, *1*, 66–80.

APA Publications and Communications Board Working Group on Journal Article Reporting Standards. (2008). Reporting standards for research in psychology: Why do we need them? What might they be? *American Psychologist*, *63*, 839–851.

Barclay, L. K. (1983). Using Spanish as the language of instruction with Mexican-American head start children: A re-evaluation using meta-analysis. *Perceptual and Motor Skills*, *56*, 359–356.

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.

Bond, Jr., C. F., Wiitala, W. L., & Richard, F. D. (2003). Meta-analysis of raw mean differences. *Psychological Methods*, *8*, 406–414.

Brown, D. (2014). *The type and linguistic foci of oral corrective feedback in the L2 classroom: A meta-analysis*. Manuscript under review.

Brutus, S., Gill, H., & Duniewicz, K. (2010). State-of-science in industrial and organizational psychology: A review of self-reported limitations. *Personnel Psychology*, *63*, 907–936.

Cameron, J., & Pierce, W. D. (1996). The debate about rewards and intrinsic motivation: Protests and accusations do not alter the results. *Review of Educational Research*, *66*, 39–51.

Chapelle, C. A., & Duff, P. A. (2003). Some guidelines for conducting quantitative and qualitative research in TESOL. *TESOL Quarterly*, *37*, 157–178.

Cheung, S. F., & Chan, D. K.-S. (2004). Dependent effect sizes in meta-analysis: Incorporating the degree of interdependence. *Journal of Applied Psychology*, *89*, 780–791.

Chiu, Y.-H. (2013). Computer-assisted second language vocabulary instruction: A meta-analysis. *British Journal of Educational Technology*, *44*, E52–E56.

Chiu, C. W. T., & Pearson, P. D. (1999, June). *Synthesizing the effects of test accommodations for special education and limited English proficiency students*. Paper presented at the National Conference on Large Scale Assessment, Snowbird, UT.

Cobb, M. (2010). *Meta-analysis of the effectiveness of task-based interaction in form-focused instruction of adult learners in foreign and second language teaching*. Unpublished doctoral dissertation, University of San Francisco.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist*, *49*, 97–1003.

Combs, J. G. (2010). Big samples and small effects: Let's not trade relevance and rigor for power. *Academy of Management Journal*, *53*, 9–14.

Cooper, H., & Hedges, L. V. (2009). Introduction. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (2nd ed., pp. 3–16). New York: Russell Sage Foundation.

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

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science*, *25*, 7–29.

DeKeyser, R., & Schoonen, R. (2007). Editors' announcement. *Language Learning*, *57*, ix–x.

Diao, N. X. (2013, March). *Cross language transfer of metalinguistic awareness: Evidence from a meta-analysis of Chinese-English bilingual children*. Paper presented at the annual conference of the American Association for Applied Linguistics, Dallas, TX.

Dollaghan, C. A., & Horner, E. A. (2013). Bilingual language assessment: A meta-analysis of diagnostic accuracy. *Journal of Speech, Language, and Hearing Research*, *54*, 1077–1088.

Ellis, N. C. (2000). Editorial statement. *Language Learning*, *50*(3), xi–xiii.

Fern, E. F., & Monroe, K. B. (1996). Effect-size estimates: Issues and problems in interpretation. *Journal of Consumer Research*, *23*, 89–105.

Fritz, C. O., Morris, P. E., & Richler, J. J. (2012). Effect size estimates: Current use, calculations, and interpretation. *Journal of Experimental Psychology: General*, *141*, 2–18.

Gass, S., Fleck, C., Leder, N., & Svetics, I. (1998). Ahistoricity revisited: Does SLA have a history? *Studies in Second Language Acquisition*, *20*, 407–421.

Gass, S., Mackey, A., & Ross-Feldman, L. (2005). Task-based interactions in classroom and laboratory settings. *Language Learning*, *55*, 575–611.

Gelman, A., & Weakliem, D. (2009). Of beauty, sex, and power: Too little attention has been paid to the statistical challenges in estimating small effects. *American Scientist*, *97*, 310–316.

Gleser, L. J., & Olkin, I. (2009). Stochastically dependent effect sizes. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis* (2nd ed., pp. 357–376). New York: Russell Sage Foundation.

Goo, J., Granena, G., Yilmaz, Y., & Novella, M. (in press). Implicit and explicit instruction in L2 learning: Norris & Ortega (2000) revisited and updated. In P. Rebuschat (Ed.), *Implicit and explicit learning of languages*. Amsterdam: John Benjamins.

Goodwin, A. P., & Ahn, S. (2010). A meta-analysis of morphological interventions: effects on literacy achievement of children with literacy difficulties. *Annals of Dyslexia*, *60*, 183–208.

Grgurović, M. (2007, October). *Research on CALL comparison studies: Can a meta-analysis inform instructed SLA*? Paper presented at the Second Language Research Forum, Urbana-Champaign, IL.

Grgurović, M., Chapelle, C. A., & Shelley, M. (2013). A meta-analysis of effectiveness studies on computer technology-supported language learning. *ReCALL*, *25*, 165–198.

Hattie, J. A. (1992). Measuring the effects of schooling. *Australian Journal of Education*, *36*, 5–13.

Hattie, J. A., Biggs, J., & Purdie, N. (1996). Effects of learning skills interventions on student learning: A meta-analysis. *Review of Educational Research*, *66*, 99–136.

Hedges, L., & Olkin, I. (1985). *Statistical methods for meta-analysis*. Orlando, FL: Academic Press.

Hill, C. J., Bloom, H. S., Black, A. R., & Lipsey, M. W. (2008). Empirical benchmarks for interpreting effect sizes in research. *Child Development Perspectives*, *2*, 172–177.

Huang, C. (2009). Magnitude of task-sampling variability in performance assessment: A meta-analysis. *Educational and Psychological Measurement*, *69*, 887–912.

Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis: Correcting error and bias in research findings* (2nd ed.). Newbury Park, CA: SAGE.

In'nami, Y., & Koizumi, R. (2009). A meta-analysis of test format effects on reading and listening test performance: Focus on multiple-choice and open-ended formats. *Language Testing*, *26*, 219–244.

Ioannidis, J. P. A. (2008). Why most discovered true associations are inflated. *Epidemiology*, *19*, 640–648.

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). Amsterdam: John Benjamins.

Jeon, E. H., & Yamashita, J. (2014). L2 reading comprehension and its correlates: A meta-analysis. *Language Learning*, *64*, 160–212.

Jun, S. W., Ramirez, G., & Cumming, A. (2010). Tutoring adolescents in literacy: A meta-analysis. *McGill Journal of Education/Revue des Sciences de L'éducation de McGill*, *45*, 219–238.

Kelley, K., & Preacher, K. J. (2012). On effect size. *Psychological Methods*, *17*, 137–152.

Kline, R. B. (2004). *Beyond significance testing: Reforming data analysis methods in behavioral research*. Washington, DC: American Psychological Association.

Kline, R. B. (2009). *Becoming a behavioral science researcher: A guide to producing research that matters*. New York: Guilford.

Koo, J., Becker, B. J., & Kim, Y.-K. (2014). Examining differential item functioning trends for English language learners in a reading test: A meta-analytical approach. *Language Testing*, *31*, 89–109.

Krashen, S. (2007). Extensive reading in English as foreign language by adolescents and young adults: A meta-analysis. *The International Journal of Foreign Language Teaching*, *3*, 23–29.

Larson-Hall, J., & Plonsky, L. (in press). Reporting and interpreting quantitative research findings: What gets reported, how, and why? In J. M. Norris, S. Ross, & R. Schoonen (Eds.), *Improving and extending quantitative reasoning in second language research*. Malden, MA: Wiley.

Lee, J., Jang, J., & Plonsky, L. (in press). The effectiveness of second language pronunciation instruction: A meta-analysis. *Applied Linguistics*.

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.

Li, S. (2010). The effectiveness of corrective feedback in SLA: A meta-analysis. *Language Learning*, *60*, 309–365.

Lipsey, M. W. (2009). Identifying interesting variables and analysis opportunities. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis* (2nd ed., pp. 147–158). New York: Russell Sage Foundation.

Lipsey, M. W., & Wilson, D. B. (1993). The efficacy of psychological, educational, and behavioral treatment: Confirmation from meta-analysis. *American Psychologist*, *48*, 1181–1209.

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.

Lyster, R., & Saito, K. (2010). Oral feedback in classroom SLA: A meta-analysis. *Studies in Second Language Acquisition*, *32*, 265–302.

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.

Masgoret, A.-M., & Gardner, R. C. (2003). Attitudes, motivation, and second language learning: A meta-analysis of studies conducted by Gardner and associates. *Language Learning*, *53*, 123–163.

Matthews, M. S., Gentry, M., McCoach, D. B., Worrell, F.C. Matthews, D., & Dixon, F. (2008). Evaluating the state of a field: Effect size reporting in gifted education. *The Journal of Experimental Education*, *77*, 55–65.

Morris, S. B., & DeShon, R. P. (2002). Combining effect-size measures in meta-analysis with repeated measures and independent-groups designs. *Psychological Methods*, *7*, 105–125.

Nakanishi, T. (2014). *A meta-analysis of extensive reading research*. Unpublished doctoral dissertation, Temple University.

Nakanishi, T. (in press). A meta-analysis of extensive reading research. *TESOL Quarterly*.

Nassaji, H. (2012). Significance tests and generalizability of research results: A case for replication. In G. Porte (Ed.), *Replication research in applied linguistics* (pp. 92–115). Cambridge, UK: Cambridge University Press.

Norris, J. M. (2012). Meta-analysis. In C. Chapelle (Ed.), *Encyclopedia of applied linguistics* (pp. 3653–3662). Malden, MA: Wiley.

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). Amsterdam: John Benjamins.

Norris, J. M., & Ortega, L. (2010). Timeline: Research synthesis. *Language Teaching*, *43*, 61–79.

Oswald, F. L., Mitchell, G., Blanton, H., Jaccard, J., & Tetlock, P. E. (2013). Predicting ethnic and racial discrimination: A meta-analysis of IAT criterion studies. *Journal of Personality and Social Psychology*, *105*, 171–192.

Oswald, F. L., & Plonsky, L. (2010). Meta-analysis in second language research: Choices and challenges. *Annual Review of Applied Linguistics*, *30*, 85–110.

Peterson, M. (2010). Computerized games and simulations in computer-assisted language learning: A meta-analysis of research. *Simulation Gaming*, *41*, 72–93.

Plonsky, L. (2009, October). *"Nix the null": Why statistical significance is overrated*. Paper presented at the Second Language Research Forum, East Lansing, MI.

Plonsky, L. (2011a). The effectiveness of second language strategy instruction: A meta-analysis. *Language Learning*, *61*, 993–1038.

Plonsky, L. (2011b). *Study quality in SLA: A cumulative and developmental assessment of designs, analyses, reporting practices, and outcomes in quantitative L2 research*. Unpublished doctoral dissertation, Michigan State University, East Lansing.

Plonsky, L. (2012a). Effect size. In P. Robinson (Ed.), *The Routledge encyclopedia of second language acquisition* (pp. 200–202). New York: Routledge.

Plonsky, L. (2012b). Replication, meta-analysis, and generalizability. In G. Porte (Ed.), *Replication research in applied linguistics* (pp. 116–132). New York: Cambridge University Press.

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., & Brown, D. (in press). Domain definition and search techniques in meta-analyses of L2 research (or why 18 meta-analyses of feedback have different results). *Second Language Research*.

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. (2012). How to do a meta-analysis. In A. Mackey & S. M. Gass (Eds.), *Research methods in second language acquisition: A practical guide* (pp. 275–295). London: Basil Blackwell.

Prentice, D. A., & Miller, D. T. (1992). When small effects are impressive. *Psychological Bulletin*, *112*, 160–164.

Reljić, G. (2011). *Does native language matter? Language achievement of preschoolers in Luxembourg and Serbia & Meta-analysis on the effectiveness of bilingual programs in Europe*. Unpublished doctoral thesis, University of Luxembourg.

Richard, F. D., Bond, C. F. Jr., & Stokes-Zoota, J. J. (2003). One hundred years of social psychology quantitatively described. *Review of General Psychology*, *7*, 331–363.

Roessingh, H. (2004). Effective high school ESL programs: A synthesis and meta-analysis. *Canadian Modern Language Review*, *60*, 611–636.

Rosenthal, R., Rosnow, R. L., & Rubin, D. B. (2000). *Contrasts and effect sizes in behavioral research: A correlational approach*. Cambridge, UK: Cambridge University Press.

Rosenthal, R., & Rubin, D. B. (1979). A note on percent variance explained as a measure of the importance of effects. *Journal of Applied Social Psychology*, *9*, 395–296.

Rosnow, R. L. & Rosenthal, R. (1989). Statistical procedures and the justification of knowledge in psychological science. *American Psychologist*, *44*, 1276–1284.

Ross, S. (1998). Self-assessment in second language testing: A meta-analysis and analysis of experiential factors. *Language Testing*, *15*, 1–20.

Rothstein, H. R., Sutton, A. J., & Borenstein, M. (Eds.). (2005). *Publication bias in meta-analysis: Prevention, assessment and adjustments*. Hoboken, NJ: Wiley.

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). Amsterdam: John Benjamins.

Sackett, P. R., & Yang, H. (2000). Correction for range restriction: An expanded typology. *Journal of Applied Psychology*, *85*, 112–118.

Schild, A. H. E., & Voracek, M. (2012). Less is less: A systematic review of graph use in meta-analysis. *Research Synthesis Methods*, *4*, 209–219.

Schwienhorst, K. (2002). The State of VR: A meta-analysis of virtual reality tools in second language acquisition. *Computer Assisted Language Learning*, *15*, 221–239.

Shintani, N., Li, S., & Ellis, R. (2013). Comprehension-based versus production-based grammar instruction: A meta-analysis of comparative studies. *Language Learning*, *63*, 296–329.

Spada, N., & Tomita, Y. (2010). Interactions between type of instruction and type of language feature: A meta-analysis. *Language Learning*, *60*, 263–308.

Stukas, A. A., & Cumming, G. (in press). Interpreting effect sizes: Towards a quantitative cumulative social psychology. *European Journal of Social Psychology*.

Sutton, A. J. (2009). Publication bias. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis* (2nd ed., pp. 435–452). New York: Russell Sage Foundation.

Tamim, R. M., Bernard, R. M., Borokhovski, E., Abrami, P. C., & Schmid, R. F. (2011). What forty years of research says about the impact of technology on learning: A second-order meta-analysis and validation study. *Review of Educational Research*, *81*, 4–28.

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. (2001). Significance, effect sizes, stepwise methods, and other issues: Strong arguments move the field. *Journal of Experimental Education*, *70*, 80–93.

Trikalinos, T. A., Churchill, R., Ferri, M., Leucht, S., Tuunainen, A., Wahlbeck, K., & Ioannidis, J. P. A. (2004). Effect sizes in cumulative meta-analyses of mental health randomized trials evolved over time. *Journal of Clinical Epidemiology*, *57*, 1124–1130.

Trikalinos, T. A., & Ioannidis, J. P. (2005). Assessing the evolution of effect sizes over time. In H. R. Rothstein, A. J. Sutton, & M. Borenstein (Eds.), *Publication bias in meta-analysis: Prevention, assessment and adjustments* (pp. 241–259). West Sussex, UK: Wiley.

Truscott, J. (2007). The effect of error correction on learners' ability to write accurately. *Journal of Second Language Writing*, *16*, 255–272.

Tukey, J. W. (1991). The philosophy of multiple comparisons. *Statistical Science*, *6*, 100–116.

Vacha-Haase, T., & Thompson, B. (2004). How to estimate and interpret various effect sizes. *Journal of Counseling Psychology*, *51*, 473–481.

Valentine, J. C., & Cooper, H. (2008). A systematic and transparent approach for assessing the methodological quality of intervention effectiveness research: The Study Design and Implementation Assessment Device (Study DIAD). *Psychological Methods*, *13*, 130–149.

Wa-Mbaleka, S. (2008). *A meta-analysis investigating the effects of reading on second language vocabulary learning*. Saarbrücken, Germany: VDM Verlag.

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.

Watanabe, Y., & Koyama, D. (2008). A meta-analysis of second language cloze testing research. *Second Language Studies*, *26*, 103–133.

Wilson, D. B., & Lipsey, M. W. (2001). The role of method in treatment effectiveness research: Evidence from meta-analysis. *Psychological Methods*, *6*, 413–442.

Won, M. (2008). *The effects of vocabulary instruction on English language learners: A meta-analysis*. Unpublished doctoral dissertation, Texas Tech University, Lubbock, TX.

Yang, J.-S. (2014, March). *The effectiveness of study-abroad in second language learning: A meta-analysis approach*. Paper presented at the annual conference of the American Association for Applied Linguistics, Portland, OR.

Yirmiya, N., Erel, O., Shaked, M., & Solomonica-Levi, D. (1998). Meta-analyses comparing theory of mind abilities of individuals with autism, individuals with mental retardation, and normally developing individuals. *Psychological Bulletin*, *124*, 283–307.

Yun, J. H. (2011a). The effects of hypertext glosses on L2 vocabulary acquisition: A meta-analysis. *Computer Assisted Language Learning*, *24*, 39–58.

Yun, J. H. (2011b). *The effects of hypertext glosses on L2 vocabulary acquisition: A meta-analysis*. Unpublished doctoral dissertation, University of Kansas.

Zhao, Y. (2003). Recent developments in technology and language learning: A literature review and meta-analysis. *CALICO Journal*, *21*, 7–27.

Supporting Information

Additional Supporting Information may be found in the online version of this article at the publisher's website:

Appendix S1: Studies included in meta-analysis