Fall 10-27-2020

A call for cautious interpretation of meta-analytic reviews

Frank Boers
fboers@uwo.ca

Lara Bryfonski

Farahnaz Faez

Todd McKay

Citation of this paper:
Boers, F., Bryfonski, L., Faez, F., & McKay, T. (2021). A call for cautious interpretation of meta-analytic reviews. *Studies in Second Language Acquisition, 43*(1), 2–24. https://doi.org/10.1017/S0272263120000327

This is a pre-final version of:
Boers, F., Bryfonski, L., Faez, F., & McKay, T. (2021). A call for cautious interpretation of meta-analytic reviews. *Studies in Second Language Acquisition*, *43*(1), 2–24.
https://doi.org/10.1017/S0272263120000327

A call for cautious interpretation of meta-analytic reviews

Abstract

Meta-analytic reviews collect available empirical studies on a specified domain and calculate the average effect of a factor. Educators as well as researchers exploring a new domain of inquiry may rely on the conclusions from meta-analytic reviews rather than reading multiple primary studies. This article calls for caution in this regard, because the outcome of a meta-analysis is determined by how effect sizes are calculated, how factors are defined, and how studies are selected for inclusion. Three recently published meta-analyses are re-examined to illustrate these issues. One illustrates the risk of conflating effect sizes from studies with different design features, another illustrates problems with delineating the variable of interest, with implications for cause-effect relations, and the third illustrates the challenge of determining the eligibility of candidate studies. Replication attempts yield outcomes that differ from the three original meta-analyses, suggesting that also conclusions drawn from meta-analyses need to be interpreted cautiously.

Keywords:

meta-analysis; replication; computer-mediated pragmatics instruction; corpus use for vocabulary learning; task-based language teaching programs.

The discipline of pedagogy-oriented applied linguistics has witnessed a proliferation of meta-analytic reviews in recent years (e.g., Lee, Jang & Plonsky, 2015; Shintani, 2015; Uchihara, Webb & Yanagisawa, 2019). These are reviews which collect as many empirical studies on the role of a given factor as possible, and then calculate the weighted average effect from that pool. Meta-analyses are useful because they help to estimate with greater confidence than any individual empirical study whether the chosen factor of interest is likely to play a role that is not confined to specific contexts, and how substantial its role is likely to be. Some researchers may therefore find meta-analytic reviews particularly useful when they make excursions into domains outside their own niche, because it seems safer to rely on a comprehensive review than on a couple of individual empirical studies. Even practitioners and policy-makers—or those advising practitioners and policy-makers—may consider the bottom line of a meta-analytic review a shortcut into the available research evidence and may rely on it to inform their instructional approaches and recommendations for teaching. Sometimes a meta-analysis may be rather broad in its research question and – though certainly of theoretical value – this may limit its potential to inform practitioners' decision making. For example, a meta-analysis which computes the likely effect of instruction in comparison with no instruction (e.g., Kang, Sok & Han, 2019) cannot, as such, tell practitioners *what* instructional interventions work particularly well, unless types of instructional interventions are examined as moderator variables as part of the analysis. In this article, however, we examine three recently published meta-analyses which are sufficiently specific in their research focus and whose conclusions may thus be taken up by educators to guide their practices. A recurring theme is the importance of cautious sampling and transparent methodological decision making, but each of the critiques serves to illustrate additional considerations for interpreting the outcomes of meta-analyses.

The first meta-analysis we examine, about pragmatics instruction (Yousefi & Nassaji, 2019), offers as one of its conclusions (regarding a moderator variable) that computer-mediated pragmatics instruction generates larger effects than face-to-face instruction. However, the collection of primary studies that this assertion is based on contains hardly any studies which directly compare the two modes of instruction. Instead, this conclusion is based on an indirect comparison of aggregated effect sizes from a small set of studies which implemented computer-mediated instruction and the aggregated effect sizes from a larger set of studies which implemented face-to-face instruction. This is potentially problematic, because effect sizes are influenced by the design features of empirical studies and by what contrasts they are based on. For example, effect sizes tend to be larger in single group pre/post study designs than in studies where effects are calculated by comparing one or more treatment groups' learning gains. A greater proportion of one type of study design is thus likely to compromise a fair comparison. A re-analysis of Yousefi and Nassaji (2019), with greater scrutiny of the primary studies and with calculation of separate effect sizes for different study designs suggests that the assertion made about the superiority of the computer-mediated mode of instruction was not (yet) justified.

The second meta-analysis we examine (Lee, Warschauer & Lee, 2019) offers strong support for the use of corpora in vocabulary learning. Unlike Yousefi and Nassaji (2019), the aggregated effect size is based exclusively on studies with a between-groups design, which should make it easier to interpret. There is nonetheless a difficulty in interpreting the outcome, because in the majority of the primary studies included in the analysis it is impossible to tell whether the between-group differences in learning gains should be ascribed to the use of a corpus *per se*, while this is the factor of interest according to the title and the abstract of the article. In some of the studies, both treatment conditions involved corpus use. In many others, the treatment conditions that involved corpus use differed from their comparison conditions in diverse ways other than corpus use. Calculating an effect size from a small set of studies where corpus use was unequivocally *the* independent variable still yields an outcome in support of corpus use for vocabulary learning, but far less compellingly so than what emerged from the original meta-analysis.

The third meta-analysis regards the benefits of task-based language teaching (TBLT) programs (Bryfonski & McKay, 2019). Although authors of primary studies often label the instructional programs they put to the test “task based” (and in this meta-analysis the same labels were used), this may not always correspond to how the approach is conceived in other TBLT literature. It is therefore difficult to determine the merits of task-*based* (versus other versions of communicative language teaching such as task-*supported* teaching) based on the aggregated effect size from the literature currently available. Replicating TBLT meta-analyses with stricter sampling criteria proves difficult because of a dearth of studies that empirically assess task-based programs relative to non-task-based programs—and the few that are available report mixed findings (e.g., Phuong, Van den Branden, Van Steendam & Sercu, 2015).

All things considered, the three “case studies” presented here illustrate that conclusions drawn from meta-analytic research should be interpreted with an eye towards the methodological choices made during the meta-analytic process.

Case Study 1: Yousefi & Nassaji (2019)

Synopsis and preliminary comments

Yousefi and Nassaji’s (2019) meta-analysis investigates the effects of instruction on second language pragmatics acquisition. According to the authors, the study is an update to prior work in this area (e.g., Jeon & Kaya, 2006, but see also another recent meta-analysis of L2 pragmatics instruction: Plonsky & Zhuang, 2019). It not only includes more recent studies but also examines previously uninvestigated moderator variables, most notably the role of computer-mediated pragmatics instruction. Based on 36 studies, the authors report overall effectiveness of pragmatics instruction as *d* = 1.101. This meta-analytic evidence that pragmatics instruction clearly works is reassuring for teachers and course designers, although instructors may be especially interested in what kinds of instruction work particularly well in certain contexts. Yousefi and Nassaji's analysis of moderator variables is informative in this regard, for example because a larger effect emerged for computer-mediated instruction (mean *d* = 1.172) than for face-to-face instruction (mean *d* = 0.965). This led the authors to assert that among the pedagogical implications of their findings "the most outstanding one is the potential of various technologies that can mediate the teaching and learning of pragmatics" (p. 25). Several additional pertinent moderator variables were explored (such as explicitness of instruction, type of outcome measures, length of treatment, and participants' proficiency level)[1], but for reasons of space, our critique will focus on the comparison of computer-mediated and face-to-face instruction.

The studies included in Yousefi and Nassaji's meta-analysis vary considerably in their designs. Many are single-group studies, where participants' progress is tracked from a pretest measure to a posttest measure (i.e., the effect size calculation is based on within-group contrasts). Others compare a treatment group's progress to that of a control group (which receives no instruction regarding the learning targets of interest in the experiment), and a few compare the progress of two treatment groups, where each group experiences a different intervention regarding the same learning targets. Yousefi and Nassaji calculated overall effects by "combining the effects of all instructional types" (p. 17). However, effects of an instructional intervention often appear larger if one contrasts participants' pre- and posttest performance than if one assesses the effectiveness of an intervention for a group of participants *relative to* another group's progress. This is because within-group comparisons of pre- and posttest scores regard the same participants in the two data sets and thus involves less variance than in the case of between-group comparisons, where the contrast in pre- to posttest gains concerns different participants (bringing in more variance). A reduction in variance and standard deviation (SD) will result in larger effect sizes because the SD makes up the denominator in the formula for Cohen's *d*. Unless studies report pre-posttest correlations which can be used as a correction for the difference, between-groups and within-groups study designs should be analyzed separately. In their meta-analysis of effects in L2 research, Plonsky and Oswald (2014) found that observed effects resulting from within-group contrasts were indeed substantially larger than between-groups contrasts. They therefore proposed a different set of benchmarks for small (d = .60), medium (d = 1.00), and large (d = 1.40) effects for within-group contrasts than for between-groups contrasts (small, d = .40, medium, d = .70 and large, d = 1.00). Owing to the mix of within-group and between-groups contrasts in Yousefi and Nassaji's collection of studies, and lack of reported pretest-posttest correlations, it is not clear how the overall estimated effect of d = 1.101 should be interpreted in relation to the above benchmarks.

We therefore re-analyzed the data by calculating separate effect sizes for the within-groups contrasts (k = 103) and the between-groups contrasts (k = 52). While we might expect such a re-analysis to produce slightly different aggregated effects sizes, we would not expect it to have profound repercussions for the general conclusion that pragmatics instruction *is* effective. As mentioned, Yousefi and Nassaji's article also investigated modality (computer-mediated versus face-to-face pragmatics instruction) as a moderator variable. An issue that can arise when examining moderators to a main effect is the difficulty in separating out and attributing unique effects to each moderating variable. In order to account for this, primary studies should be closely examined in terms of their study designs and for the potential interactions between moderating variables. For example, a recent meta-analysis about the effect of glosses on vocabulary acquisition (Yanagisawa, Webb & Uchihara, 2020) included mode of gloss (textual, pictorial or aural) as a moderating variable but deliberately selected only studies on single glosses for this comparison. Inclusion of studies on multimodal glosses would have made it difficult to separate the effect of mode (e.g., textual vs. pictorial) from the effect of providing more than one annotation (e.g., textual + pictorial) for the same word (Boers, Warren, Grimshaw & Siyanova-Chanturia, 2017; Ramezanali, Uchihara & Faez, in press).

In the case of Yousefi and Nassaji's investigation of the moderating variable of computer-mediated instruction, there is a potential interaction with the type of study design because the set of studies implementing computer-mediated instruction consists mostly of within-group contrasts, and so the larger aggregated effect size that emerged for this set could be an artefact of this design feature rather than reflecting an effect of computer-mediation *per se*. Moreover, in virtually all the computer-mediated studies the pragmatics instruction was explicit. This is relevant because Yousefi and Nassaji found a larger overall effect for explicit (d = 1.213) than implicit (d = 0.848) instructional treatments. Explicitness of instruction could thus be an alternative explanation for the comparatively large effect size that emerged from the computer-mediated interventions.

What are the contrasts?

As mentioned, there is a wide range of study designs in the collection of primary studies used by Yousefi and Nassaji (2019), yielding diverse contrasts for effect size calculations (pretest vs. posttest scores of a single group or differences in learning gains between two groups). It is important for the sake of transparency and replicability of a meta-analysis to specify what contrasts are used for these calculations (Maassen, van Assen, Nuijten, Olsson-Collentine & Wicherts, 2020). Since Yousefi and Nassaji (2019) did not include this information, we adopted the following, explicitly stated, procedures from the earlier meta-analysis by Jeon and Kaya (2006) in our replication:

1. For studies that examined one treatment group and one control group (that received no instructional intervention) by means of pre- and posttests, effect sizes were calculated by contrasting the two groups' outcomes on pre- and immediate posttests (Alcón-Soler, 2015; Bardovi-Harlig et al., 2014; Eslami & Eslami-Rasekh, 2008; Felix-Brasdefer, 2008; Furniss, 2016; Narita, 2012; Rafieyan et al., 2014; Tan & Farashaiyan, 2012).
2. For studies that examined multiple treatment groups and one control group by means of pre- and posttests, effect sizes were calculated by contrasting each group's immediate pre- and posttest outcomes separately with the control group's immediate pre- and posttest outcomes (Eslami & Liu, 2013; Hernandez, 2011; Li, 2013; Nguyen et al., 2012; Tajeddin et al., 2012).
3. For studies that examined two or more treatment groups without any control group, pretest data was contrasted with immediate posttest data for each group (Chen, 2011; Derakhshan & Eslami, 2015; Felix-Brasdefer, 2008; Fukuya & Martinez-Flor, 2008; Fordyce, 2014; Ghobadi & Fahim, 2009; Gu, 2011; Jernigan, 2012, Li, 2012a; Li, 2012b; Nguyen et al., 2015; Simin et al., 2014; Tateyama, 2007, 2009).

4. For studies that examined one group before and after an intervention, pretest data was contrasted with posttest data on immediate posttests (Alcón-Soler, 2012, Alcón-Soler & Guzman, 2010; Tanaka & Oki, 2015).

5. For studies that reported both treatment group and control group comparisons as well as within group contrasts, effect sizes were calculated for both between-group and within-group contrasts in the ways outlined above (Nguyen et al., 2012).

6. For studies that compared two groups pre- and post-intervention and only provided the results of a multifactorial test (e.g. ANOVA), the effect size was calculated from the main effect of time for each group (Takimoto, 2012a/b).

Some studies included in Yousefi and Nassaji's meta-analysis provide insufficient information to calculate effect sizes along the above procedures. It is unclear in some cases what method the original analysis utilized. For example, Dastjerdi and Farshid (2011) only reported the results of a *t*-test comparing posttest results of two experimental groups. Martinez-Flor and Alcón-Soler (2007) lacked SDs necessary to compute effect sizes (other reported statistics were nonparametric). Cunningham (2016), one of the handful of studies in the collection which implemented a computer-mediated mode of instruction, had to be excluded because the report did not provide sufficient information for calculating effects sizes comparing the two experimental groups (which only included 8 and 9 participants each). In addition, one publication (Nguyen, 2013) reported on the same data as another (Nguyen et al., 2012), and so the duplicate report was excluded. Therefore, those studies (Cunningham, 2016; Dastjerdi & Farshid, 2011; Martinez-Flor & Alcón-Soler, 2007; Nguyen, 2013) were excluded from our re-analysis leaving a total of 32 individual studies (instead of the original 36) and 155 contrasts (see supplement hosted on IRIS for a full list with justifications for inclusion/ exclusion).

Another modification to the original meta-analysis concerns the categorization of one of the studies (Nguyen et al., 2015) that was coded as computer-mediated instruction by Yousefi and Nassaji. This study utilized email writing as an outcome measure and may thus at first glance appear to be about computer-mediated instruction, but the instruction itself was not in fact computer-mediated. We therefore had to remove it from the set of computer-mediated instruction studies in our re-analysis, reducing this set to six studies.

Benefits of computer-mediated instruction?

Our re-analysis confirms the general finding of Yousefi and Nassaji (2019) that pragmatics instruction has a positive effect. For the between-groups contrasts, the overall effect is d = 1.11, a large effect for between-groups comparisons in L2 research. For within-group studies, the result is d = 1.32, a medium to large effect for within-groups comparisons (see supplement hosted on IRIS for full results tables).

However, our re-analysis does not confirm the original meta-analysis when it comes to the comparison of computer-mediated and face-to-face interventions. According to the original analysis, the former yielded larger effects (d = 1.172; k = 30[3]) than the latter (d = 0.965; k = 80), but, according to our re-analysis of the data, the face-to-face mode in fact generated the larger effects. For between-groups designs, we now find a large effect of d = 1.271 (k = 40) for face-to-face instruction and only a small effect of d = 0.65 (k = 12) for computer-mediated instruction. Taking only the within-group studies, we again find a large effect of d = 1.46 (k = 85) for face-to-face instruction and a small effect of d = .75 (k = 18) for computer-mediated instruction (see supplement hosted on IRIS for a full results table). It needs to be acknowledged that the sample sizes for the computer-mediated interventions are now even smaller than they were in the original meta-analysis (due to selection decisions explained above and due to the separation of between- and within-group contrasts). This highlights the need for more empirical investigations of computer-mediated pragmatics instruction. Investigations that *directly* compare the effectiveness of computer-mediated and face-to-face instruction for pragmatics would be especially welcome. In the collection used by Yousefi and Nassaji, only one study (Eslami & Liu, 2013) did this, and it found no difference in effectiveness between the two modes. A more recent study on pragmatics instruction (Tang, 2019), outside the scope of the meta-analysis, found no advantage for computer-mediated activities over face-to-face activities either. In sum, our replication with separate effect size calculations based on study design differences did not support the superiority of computer-mediated pragmatics instruction over face-to-face instruction.

Case Study 2: Lee, Warschauer, & Lee (2019)

Synopsis and preliminary comments

Lee et al.’s (2019) meta-analysis concerns the effects of corpus use on second language vocabulary learning. It is a partial replication of an earlier, broader-scope meta-analysis of corpus use in language learning (Boulton & Cobb, 2017) but focused specifically on vocabulary and only included studies with an instructed control group (a comparison group) in their design.[4] Based on 29 primary studies, the weighted average effect on short-term learning was found to be medium sized (Hedges’ $g = 0.74$). In eight of the studies, delayed post-tests were included, and these also showed a positive effect (Hedges’ $g = 0.64$). While Lee et al. (2019) acknowledge the role of several moderator variables (such as L2 proficiency level), the above aggregated effect sizes clearly suggest that corpus use is beneficial for L2 vocabulary learning.

Below, we highlight the issue of determining whether the main effects observed in primary studies are always a result of the variable of interest (in this case, corpus use). Before turning to that issue, we point out that it is not always clear what is meant by "effects" in this meta-analysis. Presumably, what is meant is learning outcomes. However, some of the studies (Frankenberg-Garcia, 2012, 2014; Stevens, 1991) investigated learners' success rates as they did exercises under various input conditions, but did not include posttests to gauge the learning outcomes generated by these activities.[5] If the aim of the meta-analysis was to compare the effectiveness of different procedures in terms of learning outcomes, then these studies do not serve that purpose, and so we will exclude them from our re-analysis.

What is the independent variable?

Corpora can be used for the purpose of vocabulary learning in various ways. The introduction to Lee et al. (2019) indicates that the focus of the article is corpus use for guided inductive learning (p. 722), also known as discovery learning and data-driven learning (Johns, 1991). In this approach, learners typically examine concordance lines (i.e., examples of language use extracted from a corpus) with a view to discovering the meanings of words or their usage patterns (e.g., their word partnerships or collocations). Because Lee et al. (2019) refer first (in the title and the abstract) to corpus use in general and then (in the introduction) to the benefits of concordance lines specifically for the purpose of discovery learning, there is some ambiguity about what is meant by "the effects of corpus use". If the independent variable of interest is corpus use more generally, then some of the primary studies appear not ideally suited, because both treatment conditions in these studies utilized examples extracted from a corpus. The difference between these groups was the ways in which corpus-based instances were operationalized. For example, Sun and Wang (2003) compared the use of corpus-based examples for guided inductive learning to their use for the purpose of illustrating a pattern that was first explained to the learners. In other words, it was using corpus-based instances to prompt inductive learning versus using corpus-based instances as part of deductive learning that was the variable of interest, and not the use of corpus-based instances *per se*.

If the effectiveness of *corpus use for guided inductive learning* is the main variable of interest, then the challenge is to separate the added value of corpus use from that of guided inductive learning. After all, guided inductive learning can also be steered by means of examples that are not extracted from a corpus, but that are invented or collected differently by teachers or textbook writers. With very few exceptions (e.g., Cobb, 1997; Tongpoon, 2009), the primary studies in this meta-analysis did not compare the effectiveness of corpus-based and non-corpus-based examples for the purpose of guided inductive learning. Instead, in several of the studies (Anani Sarab & Kardoust, 2014; Poole, 2012; Sripicharn, 2003; Vyatkina, 2016; Yunus & Awab, 2012) corpus-based discovery learning was compared to a condition where students received vocabulary explanations upfront followed by a few examples. In that case, it is again impossible to ascribe the superior learning observed for the corpus-based condition to the use of a corpus, because it may also be attributable to the purported benefits of guided inductive learning (as opposed to deductive learning), regardless of whether the examples used for the inductive process were extracted from a corpus or produced in another way.

There are undeniably strong arguments for the use of corpus-based examples, such as their authenticity and the ease with which many examples can be generated from an online corpus (e.g., Johns, 1991; Stevens, 1991). However, whether using corpus-based examples necessarily produces better learning outcomes than using, say, a series of textbook examples is an empirical question that is addressed by very few of the studies. Additionally, the distinction between authentic concordance lines and made-up examples can easily get blurred when researchers/ materials designers start editing concordance lines to make them more comprehensible to the learners and to ensure the discovery-learning progresses as intended (e.g., Kim, 2015; Yang, 2015). In Supatranont (2005, pp. 84-91, and appendices J and K), for example, the only difference between the concordance lines and the textbook-type examples on the student handouts was that the former looked like concordance lines while the latter were presented as regular sentences. The difference between the two treatment conditions in this study was not the presence versus absence of corpus-based examples. Nor was it the presence versus absence of discovery-learning activities, because both groups were required to find patterns in the sets of examples given on their handouts. The difference, rather, was that, in addition to pen-and-paper practice, the experimental group conducted computer-assisted searches, while the comparison group only worked with the handouts.

A level playing field?

A frequent topic in this collection of primary studies is collocation (word partnerships, such as *conduct research*, *sore throat* and *depend on*), with several studies reporting the benefits of presenting learners with sets of concordance lines showing the most common collocates of a word. The effect of exposing learners to collocations is typically shown in posttests requiring learners to recall the word partnerships they were exposed to in the treatment. However, this is often in comparison with another treatment condition which did not involve any work on collocations at all but instead included learning activities on something else, such as single words or grammar (Mirzaei, Domanaki, & Rahimi, 2016; Rahimi & Momeni, 2011; Rezaee, Marefat,

& Saeedakhtar, 2015). In other words, the experimental groups were exposed to the target items they would be tested on in the posttests, while the comparison groups were not exposed to these target items during their instructional treatment. It is therefore not surprising that the experimental groups outperformed the comparison groups in the posttests. This is reflected in some very large treatment effects (Hedges' g = 2.07 in Mirzaei, et al., 2016, and 1.98 in Rahimi and Momeni, 2011)[6]. However, whether these effects should be ascribed to the *nature* of instruction (e.g., the use of concordance lines from a corpus) or simply to the focus of instruction (i.e., collocation) is unclear. It is quite conceivable that the comparison groups would not have performed so poorly in the posttests, had they also been exposed to the target collocations during treatment. Put differently, the instructed control groups in these studies were not true comparison groups, but more akin to no-treatment control groups (i.e., groups that receive no instruction on the items or patterns that they will be tested on). If the purpose of the meta-analysis is to estimate the effectiveness of corpus use relative to other instructional treatments that share the same learning objective, then it seems justified to exclude these studies.

Other studies included in the original meta-analysis demonstrated imbalanced learning opportunities between treatment groups, even though both groups did exercises with a focus on collocation. This can be illustrated with reference to a study by Daskalovska (2012). The experimental group in this study was instructed to use online corpus tools to collect the ten most common adverb collocates of verbs and to report their findings. The comparison group did short pen-and-paper exercises about the same verbs but were exposed to a smaller number of adverbs. Obtaining a high score on one of the posttest sections—the section with the heaviest weighting—hinged on the learners' ability to supply a wide range of adverbs, and so this potentially gave an advantage to the experimental group. One of the other sections of the posttest did appear better aligned with the comparison group’s practice materials, given that it was a multiple-choice test and the study package created for the comparison group included a similar multiple-choice activity. However, the correct answers to the multiple-choice items in the posttest were *not* included in the multiple-choice exercise done in the learning stage. For example, in the exercise the students learned “I entirely agree” and “I clearly remember”, but in the post-test they needed to select “I strongly agree” and “I vividly remember” to score points. The poor posttest performance of the comparison group is therefore unsurprising.

Equally unsurprising is the finding that better learning outcomes after corpus use were observed in studies where the experimental groups engaged in corpus-based activities *in addition* to activities they shared with the comparison groups (e.g., Gordani, 2013), while comparison groups did not engage in any supplementary activities regarding the target vocabulary. In some cases, this meant the experimental groups spent extensive additional time on the target words (e.g., Karras, 2015; Yunxia, 2009). Better learning outcomes for the experimental groups in these studies could thus be attributed to differences in time investment. Supplementary activities other than corpus-based ones could also be expected to enhance learning outcomes, and so, while these studies undeniably demonstrate that corpus use is *effective*, they do not demonstrate it is *efficient* in comparison to learning activities that do not require a corpus.

There are also several publications in the collection that lack sufficient detail and transparency, and for these studies it is impossible to tell if the experimental and comparison conditions differed in more ways than use or non-use of corpus data. This lack of transparency is especially problematic given that some of these articles (some hardly four pages long) report large effects (e.g., Hedges’ g = 1.15 in Al-Mahbasi, Noor and Amir, 2015, and 1.38 in Yılmaz and Soruç, 2015), thus potentially inflating the aggregated effect.

If we re-calculate the average effect on short-term learning based on the studies from the original pool where we do feel confident enough that differences in learning outcomes can be attributed to corpus use (see supplement hosted on IRIS for the original list of studies with justification for inclusion/ exclusion), the result is markedly different from the original meta-analysis: Hedges' *g* = 0.32. According to the norms proposed by Plonsky and Oswald (2014) for between-groups contrasts, this is a small effect. However, this average is now based on only five studies, totalling only nine contrasts from the original meta-analysis. Clearly, more (and more focused) empirical investigations of the merits of corpus use are needed for a meta-analysis on this subject to produce a more reliable estimate.

Case Study 3: Bryfonski and McKay (2019)

Synopsis and preliminary comments

Bryfonski and McKay's (2019) meta-analytic review was a first effort at estimating the effectiveness of task-based language teaching (TBLT) programs.[7] Their search produced 27 studies with a between-groups design as well as a small collection of studies with within-groups designs (i.e., comparing a single treatment group's pretest and posttest performance). The original report cautioned that the number of within-groups studies was too small a collection to draw conclusions from (p. 619). Here, we therefore focus on the set of between-groups comparisons. The average effect size Bryfonski and McKay calculated from this collection (*d* = 0.93) approximates the threshold (*d* = 1.00) proposed by Plonsky and Oswald (2014) for a large between-groups effect. The report concludes that this finding "supports the notion that program-wide implementation of TBLT is effective for promoting L2 learning above and beyond the learning found in programs with other, traditional or non-task-based pedagogies" (p. 622).

One of the questions we discuss below is the extent to which the studies included in the original meta-analysis examined implementations of task-*based* language teaching, that is, TBLT in its “strong” form (Long, 2015) as opposed to task-*supported* language teaching. Before turning to this question, on reflection, it seems worthwhile to exclude three of the primary studies in the original collection of between-group studies because they examined TBLT without directly comparing TBLT to non-TBLT treatments (Lai & Lin, 2015; Li & Ni, 2013; Shabani & Ghasemi, 2014). A further study (González-Lloret & Nielson, 2015) did not establish group equivalence prior to the respective treatments (i.e., there was no pretest), and since an effect size based solely on posttest scores is not optimally reliable if we cannot be confident about pre-treatment comparability, we exclude this study in our re-analysis as well.

What’s in a name?

Some TBLT proponents distinguish between programs which use tasks throughout (Long, 2016), and task-*supported* programs, where tasks are used alongside or in addition to other approaches, including those involving explicit instruction (Ellis, 2018). With one exception (González-Lloret & Nielson, 2015—which, as already noted, was excluded from the re-analysis because of lack of pretest data), all the programs described in the primary studies included in this meta-analytic review can be considered task-supported rather than task-based. An example is Amin (2009), where “The TBL approach adopted in this study takes the form of explicit grammatical instruction in conjunction with communicative activities” (p. 81). Readers should therefore interpret TBLT, which is the term used in the majority of the included articles, as task-supported implementations, and not the “strong” version of TBLT outlined by Long (2015).

Another difficulty lies with the notion of task itself, for which slightly different definitions have been used in prior literature (e.g., Ellis, Skehan, Li, Shintani & Lambert, 2019; Long, 2015). What is agreed on by proponents of TBLT in its various forms, however, is that tasks are meaning-focused (i.e., focused on the content of messages rather than their linguistic packaging) and make learners use language as a vehicle towards a goal that itself is not linguistic. For example, in one of the original studies (Lochana & Deb, 2006) the following activities are presented as tasks according to those researchers' interpretation of TBLT: "Your teacher will read out a passage; listen to the passage carefully and complete the blanks." In another study (Amin, 2009), the author explains that "The pedagogical tasks […] are what learners do in class, such as listening to a tape and repeating phrases or sentences" (p. 44). Although these activities are labelled tasks in these publications, they are language-focused exercises rather than tasks as understood in TBLT circles. Several authors (e.g., Birjandi & Malmir, 2009; Sarani & Sahebi, 2012; Yang 2008) consider pair work as the defining characteristic of TBLT, regardless of whether the activities have a clear communicative purpose. These examples illustrate the wide interpretation of "task" in worldwide contexts.

Below we report an attempt at a new meta-analysis which adopts a narrower interpretation of tasks and which only includes studies that meet the criteria for tasks defined by Ellis and Shintani (2013, see below). First, however, it may be worth speculating why TBLT is understood in such diverse ways, including ways not at all intended by TBLT advocates. Many of the authors of the studies in the meta-analysis cited Willis (1996) and Willis & Willis (2007), summed up on http://www.willis-elt.co.uk/ and https://www.teachingenglish.org.uk/article/a-task-based-approach to justify their task and program designs. In Willis and Willis' (2007) version of TBLT, communicative tasks are preceded by a pre-task phase, to help learners prepare for the task, and are followed by a post-task phase, where time is devoted to feedback, reflection on task performance, and reactive treatment of language problems. Several authors relied heavily on this three-phase lesson model but often with a focus on language as a study object rather than as a means toward a non-linguistic end. It is understandable how "task" may be misconstrued from webpages such as https://www.teachingenglish.org.uk/article/criteria-identifying-tasks-tbl without carefully considering supplementary information. For example, one of the criteria listed there is that the activity should have "a goal or an outcome." If a researcher misinterprets this goal or outcome as increased language knowledge on the part of students, then their "TBLT" lessons may treat language as a study object instead of a vehicle. Misinformation or misunderstandings may also result in assessments of learning gains that are focused on aspects of language, such as grammatical accuracy and vocabulary knowledge, rather than the learners' successful completion of the communicative tasks (Plonsky & Kim, 2016). Once a practitioner or researcher misses the crucial point about what is meant by a goal or an outcome of a task, they may also misinterpret agreement tasks as reaching an agreement on the right answer in a language exercise and information-gap tasks as completing gap-fill exercises.

Depending on the model of TBLT, guidelines for creating task-based (or task-supported) lessons may be rather vague as to how much language-oriented instruction can (or should) be included at various stages of instruction. Additionally, TBLT proponents have slightly different views of what features distinguish a task from a language exercise. In our re-analysis, we examined the classroom procedures of the primary studies to examine the extent to which the activities labelled as tasks in the main task phase of the described lessons can be characterized as tasks as defined by Ellis and Shintani (2013). The four criteria proposed by Ellis and Shintani (2013, p. 135), slightly re-worded here, are as follows:

(1) The focus is on meaning, that is, on the content of messages rather than on the language code *per se*.

(2) There is some sort of communication gap between interlocutors, that is, learners exchange information or opinions rather than telling interlocutors—including their teacher—what these interlocutors clearly already know.

(3) The task instructions do *not* stipulate what language elements or patterns the students should use when performing the activity (because that risks turning the activity into a language-focused exercise).

(4) There should be a clear purpose (e.g., solving a problem; reaching an agreement about a dilemma) other than practicing language (because in the 'real' world, language use is a means to an end, not the end itself).

For ten of the studies, we concurred that the tasks met one out of four of criteria[8], and so it seems justified to exclude them from this narrower re-analysis (see supplement hosted on IRIS for full inclusion/ exclusion criteria). After exclusion of these and the ones mentioned in the previous section (i.e., studies which were not designed to compare task-based to non-task-based interventions), the collection includes 13 studies.

At face value?

Applying the criteria outlined above requires that authors carefully detail their instructional procedures and classroom activities. However, several of the remaining research reports provide insufficient detail to apply Ellis and Shintani's (2013) criteria. What follows are examples of how little is said about the nature of what are labelled tasks in some of the articles:

> The tasks in every lesson had a high corresponding with the course book materials, because of pre-determined syllabus. The teacher used his creativity for adaptation of the tasks with the text book. (Rezaeyan, 2014)

> [T]he students were required to do the tasks either in pair or in small groups, with the teacher monitoring their performance and encouraging more communication among them. (Mesbah, 2016)

> In task-cycle phase, the students were engaged in completing different kinds of tasks. (Tan, 2016)

> [S]tudents engaged in different communicative situations, unrelated to the actual course but organized in such a way that the participants were compelled to use the previously acquired lexico-grammar. (De Ridder et al., 2007)

> The author selected eight topics from the textbook or from outside the book, and designed the speaking tasks, considering the student's actual level and interest. (Ting, 2012)

As illustrated in the previous section, authors may cite publications about TBLT and call the classroom activities they designed tasks, but this offers no guarantee that these in fact fit the criteria for tasks established above. Some of the effect sizes in this subset of non-transparent reports are very substantial (e.g., $d > 1.7$ in Mesbah and Faghani, 2015, and in Tan, 2016), even though it is difficult to tell what these effects should be attributed to. For the sake of caution, we exclude these studies in our re-analysis as well. As a result of this, the collection now includes six studies. If these remaining studies shared a tight focus and used very similar instruments and methods, a meta-analysis of them might still be meaningful. However, they in fact display very diverse foci (e.g., speaking vs. writing skills) and outcome measures (see Saito and Plonsky, 2019, for an illustration that effect sizes can differ markedly depending on the type of outcome measures), and so it is doubtful whether a meaningful generalization can be drawn from such a small remaining sample.

A Level Playing Field?

Regardless of whether the primary studies included in the original meta-analysis really concerned TBLT programs or, instead, compared one language-focused program to another language-focused program, the fact remains that what was presented as the experimental treatment in these studies almost consistently generated the better outcomes. One might argue that, even though the experimental treatments did not meet all the criteria to be labelled task-based under our criteria, they were nonetheless better aligned with TBLT principles than the comparison treatments. If so, then the outcome of the meta-analysis could still be interpreted as support for programs exhibiting at least some features of TBLT. For example, the so-called TBLT treatments typically involved a greater amount of peer-peer interaction in the target language than the comparison treatment, where students worked mostly individually. So, even though many of the activities described in these studies are exercises instead of tasks, the fact that these exercises were typically tackled collaboratively in the treatment conditions that brought about the better learning outcomes can be meaningful (Sato & Ballinger, 2016). Put differently, more nuanced distinctions within the broad spectrum of task-*supported* programs could be fruitful to help determine the role of specific program characteristics.

As also highlighted in our discussion of Lee et al. (2019) above, better outcomes for the experimental treatment can in some cases be attributed to other factors than the so-called TBLT nature of the treatment. For example, Torky (2006) investigated the benefits of an intensive speaking course in comparison to a course where students hardly did any speaking practice. Unsurprisingly, the students from the speaking course did better in end-of-course speaking activities, which resembled their course activities. In a similar vein, the end-of-course assessment in Yang (2008) concerned speaking skills, which the experimental group had been given ample opportunity to develop in class while the comparison group had not. Considering the potential effect of practice–test congruency (i.e., the probability that one gets better at what one practices regardless of whether the practice method resembles TBLT or something else), we also exclude these studies from the collection of between-group comparisons in our re-analysis when the purpose is to gauge the effect of TBLT as an independent variable. This reduces the collection to three studies. Were we to calculate an average effect from these, the result would be d = 0.258, indicating a small effect, but this is not quite meaningful given the minute sample size.

An extra challenge with assessing many of the primary studies is that the description of the control/comparison condition[9] is often as minimal as, for example, "[the] control group experienced conventional teaching" (Rezaeyan, 2014). Even some of the lengthy texts, such as PhD dissertations, offer minimal information. For example, Murad (2009) only mentions that "the control group was taught using the conventional methods of teaching used by teachers of EFL at these schools" (p. 77), without giving any further explanations as to what those conventional methods were. When descriptions are included, these are often ambiguous as to whether the two groups spent the same amount of time on the skills or knowledge they would be needing to perform well in the post-tests. All this makes it difficult to tell whether the superior performance of the experimental group should be attributed to their being provided with *better* learning opportunities or simply *more* learning opportunities in preparation for a specific end-of-course assessment.

The latter possibility can be illustrated with two of the three studies remaining in our re-analysis. One is Lai, Zhao and Wang (2011), which did include helpful details about both the experimental and the comparison treatments as well as the assessment instruments used. In this study, communicative activities were added to a language-focused course in the experimental condition. To evaluate whether this had a positive effect on learning, a speaking test was used, where the students were asked to describe a picture of a person's bedroom (p. 96). However, picture description was a recurring course activity in the experimental condition, and one of the picture description activities in the course was about bedrooms as well (p. 102). If the students from the TBLT course performed better on the final speaking test, this may be partially attributable to practice–test congruency (because they had done the activity before while the comparison group had not). A similar example is a study by Park (2012), who designed computer-assisted activities for the TBLT group, while the non-TBLT group only worked with their prescribed EFL textbook. One of the TBLT group's computer-assisted lessons was about writing emails to e-pals (e.g., to introduce a new e-pal). The non-TBLT group, which was confined to working with the EFL textbook, appears not to have practised this specific activity. However, the same activity was used as one of the assessment measures, thus potentially giving an advantage to the TBLT group. After excluding also these two studies from the re-analysis, a single study would remain (Phuong et al., 2015). This is a study that reports a positive effect of a

TBLT-informed writing course on students' vocabulary development, but less improvement compared to the non-TBLT treatment on measures of linguistic accuracy. The result is an averaged *d*-value of -0.06. In short, using different, stricter criteria for sampling candidate studies changes the conclusions regarding the effectiveness of task-based relative to non-task-based implementations. Again, the main conclusion must be that much more (and more solid and replicable) empirical work on the comparative effectiveness of TBLT needs to be done before a robust meta-analysis of the effects of task-*based* programs will become feasible. In the interim, it is critical to apply more nuance to domain definitions within the spectrum of task-*supported* programs so that the role of specific program characteristics can be better understood.

Conclusions and recommendations

The outcome of a meta-analysis is inevitably determined by how a factor of interest is defined and how candidate studies are subsequently selected. As illustrated in all three "case studies" presented here, changes in selection criteria, such as applying more narrow definitions of key variables, can lead to different outcomes. In each of our re-analyses, we considered it desirable to exclude a fair number of studies that were included in the original meta-analyses, because they (a) were not in fact designed to address the research question that the meta-analysis sought answers to, (b) did not report quantitative data (such as pretest scores) required for a reliable effect calculation, (c) exhibited confounds that make it difficult to attribute an observed effect to the factor of interest, (d) were described with insufficient detail to allow a proper evaluation. Unfortunately, applying stricter selection criteria can drastically reduce sample sizes. If we were dealing with effect sizes from primary studies which were very precise replications of one another, then aggregated effect sizes could still be meaningful, but in the case of the three meta-analyses we have examined here we are dealing with primary studies that show considerable diversity in design, learning targets, outcome measures, and instructional settings. Given this diversity, it is not surprising that the addition or exclusion of a few primary studies can alter the outcome of a meta-analysis. The original meta-analyses seem to have been conducted in a spirit of an inclusive approach to primary study selection (for the sake of sample sizes). It has not been our intention to argue that the 'when in doubt, leave it out' stance taken in our replication attempts is necessarily better. The point is, rather, that readers of meta-analytic reviews (be they researchers, policy makers or teaching professionals) need to be aware that any meta-analytic endeavour involves multiple choices on the part of the analyst, each of which impacts the outcomes (Oswald & Plonsky, 2010). To help readers appreciate this, authors of meta-analytic reviews are of course urged to be totally transparent about the choices they made (Maassen et al., 2020; Norris & Ortega, 2006). It is doubtful, however, whether many consumers of meta-analytic reviews closely inspect the method sections in such publications, where those choices are explained. Instead, readers may rely solely on the information provided in the abstract and possibly the general conclusion section. Owing to their status as comprehensive reviews, conclusions drawn from meta-analyses exert a certain authority. We hope to have demonstrated that assertions about the role of a given factor (be it the primary factor of interest or a moderating factor) need to be made with caution, especially in the case of recent strands of empirical inquiry.

Recommendations may also be distilled for the researcher wishing to embark on a meta-analysis. One recommendation is to carefully delineate the factor(s) of interest and to evaluate whether the available strand of research related to this factor lends itself to a robust and meaningful analysis. When the maturity of a given domain for meta-analysis is uncertain, it is recommended to first carry out a scoping review. A scoping review is another type of research synthesis that surveys a domain of literature identifying current trends, commonly used methods, and gaps in findings (e.g., Gurzynski-Weiss & Plonsky, 2017; Hillman, Selvi & Yazan, in press; Tullock, & Ortega, 2017). A scoping review can help determine if subsequent meta-analytic work is appropriate and worthwhile. After embarking on a meta-analysis, researchers are advised to scrutinize each candidate study to determine its eligibility and make the criteria for study inclusion clear. As we have illustrated, a field may look ready at first glance, as one starts deploying the powerful online search engines at our disposal, but this may be deceptive if it turns out that many candidate studies fail to meet the standards for inclusion. Unfortunately, scrutinizing the method sections of a large collection of empirical research papers is a labour-intensive exercise. Meta-analytic replications are of course not immune to interpretation errors either. We fully recognize potential shortcomings in our own reassessment of the primary studies included in our three case studies. Alternatively, a faster way could be to use the prestige of the journals where they were accepted as a proxy of quality assuredness (e.g., Faez, Karas & Uchihara, in press), under the assumption that some journals use more rigorous review processes than others. This, then, raises the difficult question what bibliometric data are most suitable to distinguish between journals on account of the relative rigour of their review processes. An additional difficulty is that resulting literature from this approach may be limited to publications from privileged, "WEIRD" (Western, Educated, Industrialized, Rich and Democratic) contexts, potentially disadvantaging those who have less access to publishing in prestigious peer-refereed journals (Andringa & Godfroid, in press; Cho, 2009; Henrich, Heine & Norenzayan, 2010). Besides, even the most prestigious journals occasionally publish articles that are arguably non-optimal (or, at least, non-optimally suited for a given meta-analytic purpose). In fact, among the primary studies we felt it justified to exclude from our re-analyses, there were indeed several ones which appeared in prestigious journals[10] (See supplements on IRIS for details on each individual study). It is worth mentioning in this context that each of the three meta-analytic reviews examined here appeared in prestigious journals, too. So, perhaps our call for caution should be extended to journal editors, editorial boards, and reviewers.

In any case, given the issues highlighted, (some of) the conclusions presented in the meta-analyses we have examined here should be taken as tentative for now. Fortunately, as new studies are continually being added to the various strands of inquiry in our discipline, we must be hopeful that sooner or later it will become possible to revisit these meta-analyses and to replicate them with a larger collection of eligible studies. This sustained effort at updating and replicating meta-analyses can be made lighter if meta-analytic reports themselves are transparent not only as to what studies were included but also as to precisely how effect sizes were calculated (so the same procedures may be followed in the updates). For one of the three meta-analyses examined here (Yousefi & Nassaji, 2019), we felt it necessary to re-calculate effect sizes because it was not clear to us precisely what contrasts the authors had based their calculations on. A lack of clarity of how contrasts were defined and analyzed not only limit readers' ability to evaluate meta-analytic findings, but it also hinders replication where effect sizes from new studies could systematically be added to an existing pool and thus gradually make the outcome more robust. The field of applied linguistics has heralded a push towards open-science practices in recent years, including recognition of open data and materials through badges in major journals (e.g., *Studies in Second Language Acquisition, Annual Review of Applied Linguistics, Language Learning, Modern Language Journal*), repositories for instruments and materials (IRIS-database.org), and registered replications (Morgan-Short et al., 2018) and reports (Marsden et al., 2018). Open science practices are one way to promote equity through the sharing of knowledge, instruments and findings in freely accessible and permanent repositories. While there is growing excitement around open access in applied linguistics research, L2 researchers (and academics more broadly) often fail to practice what they preach in terms of publishing open-access (e.g., Zhu, 2017) or to making data freely available. The coding schemes and data of some prior meta-analyses have been uploaded in repositories such as IRIS (Bryfonski & McKay, 2017; Plonsky, 2011, 2012, 2019; Plonsky & Kim, 2016; Plonsky & Ziegler, 2016; Plonsky & Zhuang, 2019), and this is also where the coding schemes and data of the present three replications can be found. Others have called for more attention to open science in meta-analytic work; McKay and Plonsky (in press), for example, recommend that "all meta-analysts make available not only their coding schemes but also their data and any code used to analyze that data" (p. 14). However, meta-analysis continues to be under-represented in terms of shared materials and data. Open data is yet another methodological choice, one that may open the door more easily to scrutiny of studies and findings. Whatever channel is deemed most appropriate, the sharing of coding sheets in meta-analysis is critical for building upon prior work and supporting future meta-analysts. It is worth mentioning that calls for greater transparency in reporting meta-analyses are being made outside the discipline of Applied Linguistics as well (e.g., Maassen et al, 2020, in the field of psychology).

Returning specifically to the three case studies we have presented here, it is important to clarify that our intention was by no means to criticize the instructional interventions advocated in them (i.e., technology-mediated pragmatics instruction, corpus use for vocabulary learning, and task-based language teaching). It was, in fact, our interest in these topics which led us to read and then further explore these three meta-analyses. We hope that our three examples can serve as an incentive for others to re-examine the meta-analyses available in their own domains of interest.

Notes

1. Despite the title of Yousefi and Nassaji's (2019) article, "A meta-analysis of the effects of instruction and corrective feedback on L2 pragmatics and the role of moderator variables", the effect of corrective feedback is not investigated in their analysis. This is surprising also because the article appeared in a special issue on the theme of "technology-mediated feedback and instruction." It is possible that the authors prioritized the topic of technology in their analysis, which would then also explain their foregrounding of the potential of computer-mediated instruction.

2. For studies that did not report pre- and posttest correlation, a conservative estimate of .30 was utilized during effect size calculation.

3. There is some inconsistency in Yousefi and Nassaji's article as regards the number of unique samples included in their calculations. It is first said that (after removing outliers) there were 27 computer-mediated and 83 face-to-face samples, but the results table later mentions 30 and 80, respectively.

4. Although Lee et al. (2019) intended to include only studies with an instructed control group (or comparison group), we failed to find information about such a group in one of the publications. This is an article (Horst, Cobb & Nicolae, 2005) that describes the design and development of a module of computer-assisted corpus-based activities. The module was tried at different stages of development with different cohorts of students, but we found no mention of a non-corpus comparison treatment.

5. There is an additional study that investigated how much students were helped by certain resources as they tackled vocabulary exercises. Kaur and Hegelheimer (2005) examined

students' success rates on vocabulary exercises either with the assistance of both an online dictionary and a concordancer or with the assistance of the online dictionary only. To estimate learning outcomes, the students' voluntary use of target vocabulary in an essay they wrote outside of class was assessed. Inter-rater reliability was only .68, however. No pretest data are included, which also makes it hard to assess learning gains as a result of the exercises, and so it felt prudent to exclude this study as well.

6. The effect sizes we mention in this section are the ones calculated by Lee et al. (2019; online supplement).

7. Other meta-analytic reviews on the subject of TBLT are available (e.g., Cobb, 2010), but these do not focus specifically on task implementation over an extensive period of time (such as a complete school term), while the subject of Bryfonski and McKay (2019) is TBLT *program* implementations.

8. The one criterion met in these studies was the meaning-focused nature of the activities, for example because they focused on text comprehension. The criterion met the least often in the collection of primary studies was having a clear non-linguistic purpose for doing the activity.

9. Most of the primary studies use the term control group in the sense of comparison group (i.e., not in the sense of no-treatment group).

10. For example, one of the publications we have had to exclude (De Ridder et al., 2007) when revisiting Bryfonski and McKay's (2019) collection of studies, was a brief report in the Forum section of the prestigious journal *Applied Linguistics*. It was felt necessary to exclude it because (a) the description of the task-based component of the course was too vague to meet the stricter sampling criteria and (b) the end-of-course assessment was different for the experimental and the comparison group, thus introducing a confounding variable.

References

(Note: Studies from the original meta-analysis can be found with all supplementary material on IRIS).

Andringa, S. & Godfroid, A. (in press). On the foundations of knowledge in Applied Linguistics research: Sampling bias and the problem of generalizability. *Annual Review of Applied Linguistics*, 40.

Boers, F., Warren, P., Grimshaw, G., & Siyanova-Chanturia, A. (2017). On the benefits of multimodal annotations for vocabulary uptake from reading. *Computer Assisted Language Learning, 30*, 709–725. https://doi.org/10.1080/09588221.2017.1356335

Boulton, A., & Cobb, T. (2017). Corpus use in language learning: A meta-analysis. *Language Learning, 67*, 348–393. https://doi.org/10.1111/lang.12224

Bryfonski, L., & McKay, T. H. (2019). TBLT implementation and evaluation: A meta-analysis. *Language Teaching Research, 23*, 603–632. https://doi.org/10.1177/1362168817744389

Cho, D. W. (2009). Science journal paper writing in an EFL context: The case of Korea. *English for Specific Purposes*, 28(4), 230-239. https://doi.org/10.1016/j.esp.2009.06.002

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. PhD dissertation, University of San Francisco.

Ellis, R. (2018). *Reflections on task-based language teaching*. Bristol: Multilingual Matters. https://doi.org/10.21832/9781788920148

Ellis, R., & Shintani, N. (2013). *Exploring language pedagogy through second language acquisition research*. New York: Routledge. https://doi.org/10.4324/9780203796580

Ellis, R., Skehan, P., Li, S., Shintani, N., & Lambert, C. (2019). *Task-based language teaching: Theory and practice*. Cambridge: Cambridge University Press. https://doi.org/10.1017/9781108643689

Faez, F., Karas, M. & Uchihara, T. (in press). Connecting language proficiency to teaching ability: a meta-analysis. *Language Teaching Research.* https://doi.org/10.1177/1362168819868667

Gurzynski-Weiss, L., & Plonsky, L. (2017). Look who’s interacting: A scoping review of research involving non-teacher/non-peer interlocutors. In L. Gurzynski-Weiss (Ed.), *Expanding individual difference research in the interaction approach: Investigating learners, instructors, and other interlocutors* (pp. 305–324). Philadelphia, PA: John Benjamins.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The Weirdest People in the World? *Behavioral and brain sciences, 33,* 61–83. https://doi.org/10.1017/S0140525X0999152X

Hillman, S, Selvi, A. F., Yazan, B. (in press). A scoping review of world Englishes in the Middle East and North Africa. *World Englishes*. https://doi.org/10.1111/weng.12505

Jeon, E. H., & Kaya, T. (2006). Effects of L2 instruction on interlanguage pragmatic development. In J. M. Norris & L. Ortega (Eds.) *Synthesizing research on language learning and teaching* (pp. 165–211). John Benjamins Publishing. https://doi.org/10.1075/lllt.13.10jeo

Johns, T. (1991). Should you be persuaded: Two examples of data-driven learning. *English Language Research Journal, 4,* 1–16.

Kang, E. Y., Sok, S., & Han, Z. (2019). Thirty-five years of ISLA on form-focused instruction: A meta-analysis. *Language Teaching Research, 23*, 428–453. https://doi.org/10.1177/1362168818776671

Lee, H., Warschauer, M., & Lee, J. H. (2019). The effects of corpus use on second language vocabulary learning: A multilevel meta-analysis. *Applied Linguistics, 40,* 721–753. https://doi.org/10.1093/applin/amy012

Lee, J., Jang, J., & Plonsky, L. (2015). The effectiveness of second language pronunciation instruction: A meta-analysis. *Applied Linguistics*, *36*, 345–366. https://doi.org/10.1093/applin/amu040

Long, M. (2015). *Second language acquisition and task-based language teaching*. Oxford: Wiley-Blackwell.

Long, M. (2016). In defense of tasks and TBLT: Nonissues and real issues. *Annual Review of Applied Linguistics*, 36, 5–33. https://doi.org/10.1017/S0267190515000057

Maassen, E., van Assen, M. A. L. M., Nuijten, M. B., Olsson-Collentine, A., Wicherts, J. M. (2020). Reproducibility of individual effect sizes in meta-analyses in psychology. *PLoS One, 15*, e0233107. https://doi.org/10.1371/journal.pone.0233107

Marsden, E., Morgan-Short, K., Trofimovich, P., & Ellis, N. C. (2018). Introducing Registered Reports at Language Learning: Promoting Transparency, Replication, and a Synthetic Ethic in the Language Sciences. *Language Learning*, *68*, 309–320. https://doi.org/10.1111/lang.12284

McKay, T. H., & Plonsky, L. (in press). Reliability analyses: Evaluating error. In P. Winke and T. Brunfaut (Eds.*), The Routledge handbook of second language acquisition and language testing*. New York: Routledge.

Morgan-Short, K., Marsden, E., Heil, J., Ii, B. I. I., Leow, R. P., Mikhaylova, A., … Szudarski, P. (2018). Multisite replication in second language acquisition research: Attention to form during listening and reading comprehension. *Language Learning*, *68*, 392–437. https://doi.org/10.1111/lang.12292

Norris, J. M., & Ortega, L. (2000). Effectiveness of L2 instruction: A research synthesis and quantitative meta-analysis. *Language learning*, *50*, 417–528. https://doi.org/10.1111/0023-8333.00136

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. 1-50). John Benjamins.

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

Plonsky, L. (2011). The effectiveness of second language strategy instruction: a meta-analysis. *Language Learning, 61*, 993–1038. https://doi.org/10.1111/j.1467-9922.2011.00663.x

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

Plonsky, L. (2019). Recent research on language learning strategy instruction. In A. Chamot & V. Harris (Eds.), *Learning strategy instruction in the language classroom: Issues and implementation*. Bristol, UK: Multilingual Matters.

Plonsky, L., & Kim, Y. (2016). Task-based learner production: A substantive and methodological review. *Annual Review of Applied Linguistics, 36*, 73–97. https://doi.org/10.1017/S0267190516000015

Plonsky, L. & F. L. Oswald (2014). How big is 'big'? Interpreting effect sizes in L2 research. *Language Learning, 64*, 878–912. https://doi.org/10.1111/lang.12079

Plonsky, L., & Oswald, F. L. (2015). Meta-analyzing second language research. In L. Plonsky (ed). Advancing quantitative methods in second language research (pp. 106–128). New York: Routledge. https://doi.org/10.4324/9781315870908-6

Plonsky, L., & Zhuang, J. (2019). A meta-analysis of second language pragmatics instruction. In N. Taguchi (Ed.), *The Routledge handbook of SLA and pragmatics* (pp. 287–307). New York: Routledge.

Plonsky, L., & Ziegler, N. (2016). The CALL-SLA interface: Insights from a second-order synthesis. *Language Learning and Technology, 20*, 17–37.

Ramezanali, N, Uchihara, T., & Faez, F. (in press). Efficacy of multimodal glossing on second language vocabulary learning: A meta-analysis. *TESOL Quarterly*. https://doi.org/10.1002/tesq.579

Sato, M. & S. Ballinger (Eds.) (2016). *Peer interaction and second language learning: Pedagogical potential and research agenda*. Amsterdam: John Benjamins. https://doi.org/10.1075/lllt.45

Saito, K., & Plonsky, L. (2019). Measuring the effects of second language pronunciation teaching: A proposed framework and meta-analysis. *Language Learning, 69*, 652–708. https://doi.org/10.1111/lang.12345

Shintani, N. (2015). The effectiveness of Processing Instruction and Production-based Instruction on L2 grammar acquisition: A meta-analysis. *Applied Linguistics, 36*, 306–325. https://doi.org/10.1093/applin/amu067

Tang, X. (2019). The effects of task modality on L2 Chinese learners' pragmatic development: Computer-mediated written chat vs. face-to-face oral chat. *System, 80*, 48–59. https://doi.org/10.1016/j.system.2018.10.011

Tullock, B., & Ortega, L. (2017). Fluency and multilingualism in study abroad: Lessons from a scoping review. *System, 71*, 7–21. https://doi.org/10.1016/j.system.2017.09.019

Uchihara, T., S. Webb, S., & Yanagisawa, A. (2019). The effects of repetition on incidental vocabulary learning: A meta-analysis of correlational studies, *Language Learning, 69*, 559–599. https://doi.org/10.1111/lang.12343

Willis, J. (1996). *A framework for task-based learning*. Essex: Longman.

Willis, D. & J. Willis (2007). *Doing task-based teaching*. Oxford: Oxford University Press.

Yanagisawa, A., Webb, S., & Uchihara, T. (2020). How do different forms of glossing contribute to L2 vocabulary learning from reading? A meta-regression analysis. *Studies in Second Language Acquisition, 42*, 411–438. https://doi.org/10.1017/S0272263119000688

Yousefi, M., & Nassaji, H. (2019). A meta-analysis of the effects of instruction and corrective feedback on L2 pragmatics and the role of moderator variables: Face-to-face vs. computer-mediated instruction. *ITL-International Journal of Applied Linguistics, 170*, 277–308. https://doi.org/10.1075/itl.19012.you

Zhu, Y. (2017). Who support open access publishing? Gender, discipline, seniority and other factors associated with academics' OA practice. *Scientometrics, 111*, 557–579. https://doi.org/10.1007/s11192-017-2316-z