Test-Practice Effects in L2 Grammar Instruction Research: A Pilot Meta-Analysis

Ryo Maie

Second Language Studies, Michigan State University

maieryo@msu.edu

Abstract

One experimental phenomenon that can cause a threat to the internal validity of L2 teaching research is *test-practice effects*: the extent to which L2 learners improve by taking the same or similar tests for multiple occasions (e.g., at a pretest and posttests). This article reports a pilot meta-analysis within a larger study that investigated test-practice effects in L2 grammar instruction research. Specifically, I focused on how the test-only control group (i.e., participants who only receive tests without instructional treatments) generally improve from the pretest to the immediate posttest and the delayed posttest. I retrieved and coded 21 primary studies of L2 grammar instruction research, which included 23 independent samples contributing 100 effect sizes in the form of Cohen's *d*. Using multilevel random-effects models, I estimated the overall population estimate of the effect sizes and investigated whether two moderator variables, the mode of language use (comprehension or production) and the type of outcome measures (explicit or implicit measures), changed the size of the test-practice effects. Results showed that test-practice effects indeed exist in L2 grammar instruction research, and their sizes lie between small and medium effects (d = 0.116−0.315). The effects also became more prominent when the test drew upon comprehension skills (d = 0.179−0.466) or used implicit measures of L2 knowledge (d = 0.173−0.597).

Keywords: testing learning effects, instructional efficacy, L2 instruction, meta-analysis

Introduction

In second language (L2) instruction research, researchers investigate "how the systematic manipulation of the mechanisms of learning and/or the conditions … enable or facilitate the development of a language other than one's first" (Loewen, 2020, p. 2). In classroom teaching, systematic manipulation typically takes the shape of instructional interventions by which a teacher intervenes in the learner's learning processes to promote L2 development. One type of instructional interventions that has been the target of extensive research is form-focused instruction (FFI), defined by Ellis (2001, p. 2) as "any planned or incidental instructional activity that is intended to induce language learners to pay attention to linguistic form." Although FFI is an umbrella term (see Loewen, 2020, Ch.4 for a review), incorporating many different teaching techniques, popular inquiries of research include the role of explicit and implicit instruction (Goo et al., 2015; Norris & Ortega, 2000; Spada & Tomita, 2010) and oral (Li, 2010; Lyster & Saito, 2010; Russel & Spada, 2006) and written corrective feedback (Kao, 2013; Kang & Han, 2015; Lim & Renandya, 2020). A series of meta-analyses provide consistent evidence that these interventions have medium to large effects on L2 learning (or part thereof) when learning is defined as (any) improvement from a pretest to posttest(s). However, when L2 instruction research adopts the pre-posttest design, a study may suffer from the problem of *test-practice effects*, that is, the extent to which learners improve by taking the same or similar tests for multiple occasions. If a researcher's interest is to examine how effective a given instructional intervention is, test-practice effects can threaten the internal validity of the study because it makes it unclear whether and to what extent learning gains from a pretest to a posttest are truly attributable to the instruction (see the next section for details). In this article, I have reported results of a pilot study for a larger research program in which the overall size of test-practice effects is examined among L2 studies that investigated efficacy of various instructional interventions on learning L2 grammar. In the following, I will briefly describe the context of the pilot study and then move onto more specific aspects of the study, including research questions, methods, and results.

Experimental Designs and Test-Practice Effects in L2 Instruction Research

Questions about the efficacy of L2 instruction can take many different forms, but by and large, they mainly address two types of instructional efficacy: (a) the absolute effect and (b) the relative utility (see Doughty, 2003; Long, 1983, 1988 for detailed reviews). The absolute effect refers to the degree of L2 learning enabled or facilitated by instruction but nothing else, whereas the relative utility is the comparative effectiveness of an instructional treatment compared to some other method of teaching. An answer to each efficacy type involves demonstrating a causal relationship between L2 instruction on the one hand and rate or ultimate attainment in L2 development on the other (Long, 1983, 1988), but each type of instructional efficacy requires different experimental designs so that such causal relationship is ascertained. Figure 1 depicts typical experimental designs of L2 instruction research for illustration. Because an absolute effect of instruction is shown when (and only when) changes in rate or ultimate attainment in L2 development are attributed to instructional treatments *only*, evidence needs to come through the comparison between an experimental group and a control group on their gains from a pretest to posttests. Control group in this respect means a true, test-only control group that does not receive any instructional intervention.[1] In Figure 1, this comparison corresponds to the contrast between an experiment group (Experimental A or B) and a control group (Control). The relative utility, on the other hand, can be shown in two ways: by comparing two different experimental groups (Experiment A vs. B) or by comparing an experimental group with a comparison group. A comparison group is a counterfactual group that encodes the status quo in teaching and is hence taught by a method that is widely conducted but considered not ideal. In L2 instruction research, a comparison group is often recruited as an intact L2 class that is taught with the same method that has been used in the class. The goal here is to investigate how a new teaching method in the experimental group improves upon the status quo of the comparison group. Although both types of instructional efficacy present useful information in favor (or disfavor) of the instructional intervention of one's interest, the absolute effect has the advantage of greater generalizability because its interpretation is not dependent on a specific contrast with a particular experimental group or a comparison group (that is taught by a specific teaching method).

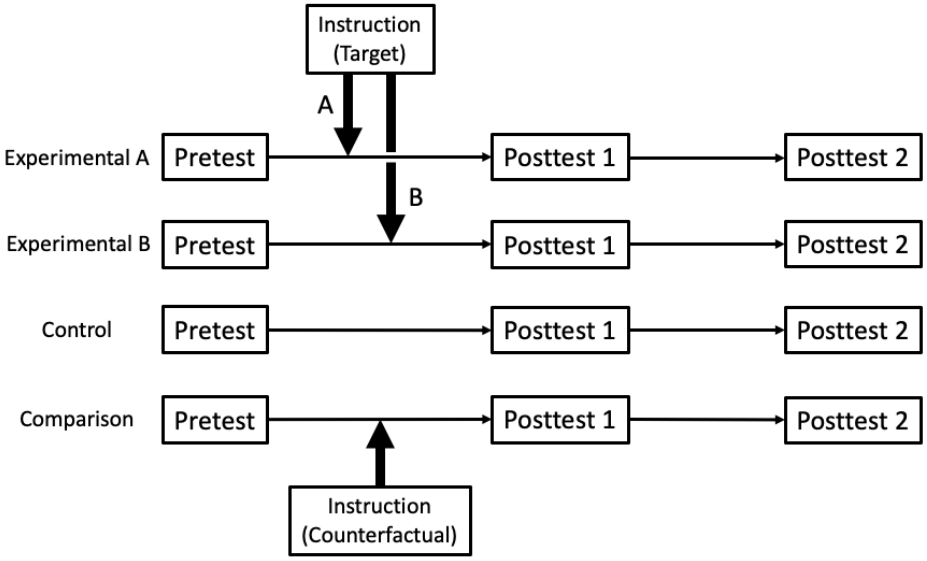

Figure 1. Typical designs of L2 instruction research.

One experimental phenomenon that is known to cause a threat to the internal validity (i.e., the causal relationship between L2 teaching and learning) of L2 instruction research is test-practice effects, i.e., the extent to which learners improve by taking the same or similar tests for multiple occasions.[2] As research questions asked in L2 instruction research typically require a pretest-posttest(s) design, test-practice effects are often inevitable. For instance, suppose that a researcher investigated the effectiveness of classroom activity in which learners orally practiced producing the subject relative pronoun in English (e.g., *The man who is watching the bird is Shawn*). The researcher assessed the learners' gains using an elicited imitation test taken as a pretest and an immediate posttest. When one finds a meaningful difference (in any form) between the two testing points, how much of the improvement can be attributed to: (a) the classroom activity, (b) taking the test (with multiple items) more than once, or (c) the combination of both? The only way to assess (but not circumvent) the degree of test-practice effects is to recruit a test-only control group to examine how much control group participants improve from the pretest to the posttest. By subtracting the gain of the control group from that of the experimental group, one can presumably identify to what degree learning gains can be genuinely attributable to instruction. Unfortunately, the use of a control group is not highly prevalent in L2 instruction research. For instance, only 23% of the studies in the meta-analysis by Norris and Ortega (2000) included a control group (of any kind); and after 15 years, Goo et al. (2015) reported that only 59% of the studies recruited a control group.

Surely, the seriousness of test-practice effects can depend on a study's research questions. When a researcher investigates the relative utility of two (or more) instructional interventions, test-practice effects may not be of much concern because two (or more) groups are compared across the board, including the effects of both teaching and testing combined. This means that the effect of teaching does not need to be isolated from that of testing. However, findings on relative utility may not be as generalizable as one would expect because interpretation is contingent on the comparison of specific teaching methods. In this respect, the absolute effect fares a better chance of generalizing to wider teaching contexts, but its evidence, unlike that of relative utility, can be (significantly) contaminated by test-practice effects. Again, using a test-only group can help assess the degree of test-practice effects, but such practice is not widespread in L2 instruction research (see e.g., Goo et al., 2015; Norris & Ortega, 2000).

As the only empirical study of test-practice effects in L2 instruction research, Suga (2022) investigated how L2 learners of English improved their skills of using target grammar structures solely by taking an elicited imitation test three times. Suga pointed out that when one investigates the effectiveness of various types of FFI (e.g., corrective feedback), using an elicited imitation test as an outcome measure can lead to a strong test-learning effect because taking the elicited imitation test itself can serve as additional practice opportunities. This point is well taken, given that existing FFI studies often adopt more than a dozen of test items for an elicited imitation test (e.g., n = 12 in Ellis et al., 2006; n = 30 in Li et al., 2016; n = 32 in Loewen et al., 2009). For instance, in Loewen et al. (2009), learners' gains (on learning the third-person -*s* in English) from a pretest to an immediate posttest and a delayed posttest was measured using an elicited imitation test with 32 items. This meant that across the study, the learners had a total of 96 trials to practice the target structure besides instructional treatments. In Suga (2022), two groups of learners with differing numbers of test items (n = 8 and 16 per structure) significantly improved from a pretest to an immediate posttest (d = 0.47−0.71) and to a delayed posttest (d = 0.51−0.82) even though the learners did not receive any instructional treatment on the target structures. Furthermore, the test-practice effect tended to be more severe when the learners received 16 test items than 8 items.

Although previous FFI studies consistently reported the existence of test-practice effects (see Ellis et al., 2005; Li et al., 2016; Loewen et al., 2009), Suga (2022) is the only study that empirically investigated the issue in L2 instruction research. One alternative method to gauge test-practice effects is to review how a control group in previous studies improved from a pretest to posttests and statistically summarize the effect using a meta-analytic technique. The current pilot meta-analysis followed the same spirit as Suga (2022) in order to assess the degree of test-practice effects in L2 grammar instruction research. While various types of linguistic features can be the target of instruction, grammar (i.e., morphology and syntax) has been the most popular target area (see Sok et al., 2019). As not many studies in L2 instruction research make use of a test-only control group (Goo et al., 2015; Norris & Ortega, 2000), synthesizing test-practice effects across the domain of L2 grammar instruction research will not only reveal whether such empirical phenomenon indeed exists, but also help us re-interpret the effectiveness of L2 instruction shown in previous studies that did not recruit a control group. While Suga (2022) investigated how test-practice effects may change depending on the number of test items (8 or 16) and the type of target features (syntactic or morphological), there are many other variables that can potentially influence the size of test-practice effects (see Conclusions and Discussions for a list). In addition to meta-analyzing test-practice effects in L2 grammar instruction research, I investigated (a) the mode of language use or skills (comprehension vs. production) and (b) the type of outcome measures (explicit vs. implicit measures) as moderating variables.

The Study

The study reported here is a pilot study of an ongoing research meta-analysis project that investigates test-practice effects in L2 grammar instruction research. Although the scope of the main project encompasses both control and comparison groups, the current pilot study only focused on test-practice effects in the control group, i.e., a group of participants that only received a pretest and posttests and did not experience any instruction intervention. The following research questions guided the study:

1. When investigating the effects of instruction on L2 grammar, what is the overall size of test-practice effects, that is, the degree to which L2 learners improve due to taking the same or similar tests for multiple occasions?
2. To what extent do two moderating variables, the mode of language use (comprehension vs. production) and the type of outcome measures (explicit vs. implicit measures), change the size of the test-practice effect?

In the following sections, I have addressed the two research questions by quantitatively synthesizing effect sizes from primary studies of L2 grammar instruction research. Specifically, I estimated the overall population-level effect size of the control group's improvement from the pretest to the posttests (Research Question 1) and examined whether the two variables moderated the size of the overall test-practice effect (Research Question 2). The current study utilized the standard procedure for typical meta-analytic studies to retrieve, code, and analyze primary studies (Cooper et al., 2009; see Plonsky & Oswald, 2015 for a review in L2 contexts).

Methods

Study Retrieval

The first procedural step in conducting a meta-analysis is to decide on how to identify a body of literature. For this pilot study, I conducted an expedited search in which I first searched two reference databases, *Linguistics and Language Behavior Abstract* and *Educational Resources Information Center*, for previous meta-analyses of L2 grammar instruction research, and only retrieved primary studies that were already included in the previous meta-analyses. Although a researcher conducting a meta-analysis typically searches primary studies, this expedited search allowed me to identify the existing literature rather quickly, which was deemed sufficient and satisfactory for the current purpose: a preliminary investigation of test-practice effects in L2 grammar instruction research. The following combinations of keywords were used to search the databases (abstract search):

> (second language OR foreign language OR L2 OR FL) AND (instruction OR teaching OR treatment OR focus on form OR form-focused OR feedback) AND (meta-analysis OR meta-analytic)

This step resulted in 17 meta-analytic reports on the effectiveness of L2 grammar instruction. I then inspected each primary study included in the meta-analyses and checked whether the study met the following inclusion/exclusion criteria:

1. The study adopted a pretest-posttest(s) design.
2. The study made use of a test-only control group.
3. The study investigated the effects of instructional interventions on L2 grammar, including morphology, syntax, and/or morphosyntax.
4. The study reported both the mean and standard deviation of learners' performances at the pretest and the posttests.

Since the primary studies rarely reported effect sizes (of improvement from the pretest to the posttests) for a control group, I instead looked for the mean and standard deviation of learners' performances based on which the target effect size (i.e., Cohen's *d*) was calculated. For this reason, those studies that failed to report the two measures of central tendency were excluded from the analysis. After filtering out primary studies that did not meet the four criteria, the final dataset consisted of 21 studies with 23 independent control groups (*n* = 393 participants), contributing 100 effect sizes in total (but see below because each effect size was calculated at five levels of correlation between the pretest and the posttests). The mean number of participants across the control groups was 17.08 participants (*SD* = 10.18; 95% CI [12.88, 21.29]), with the minimum and maximum number being 6 and 45 participants, respectively. Note that when a study comprised of more than one experiment with different participant samples, I coded each as a separate study. Furthermore, if a study is part of a larger study, I only included the latter.

Coding

Each primary study was coded for 21 study characteristics, summarized in Table 1 below. The target effect size was Cohen's *d*, a metric that quantifies the difference between two means on a standardized scale (i.e., the standard deviation unit). In the dataset, 15 studies (71.42%) administered posttests at more than one time point (i.e., immediate and delayed posttests), whereas only one (4.76%) did so at more than two time points (i.e., immediate posttest and more than one delayed posttest). Thus, I decided to focus on the difference between the pretest and the immediate posttest (Pretest vs. Immediate), and the difference between the pretest and the first delayed posttest (Pretest vs. Delayed). Any test sessions after the first delayed posttest were not analyzed. Because the studies rarely reported effect sizes for a control group, I derived Cohen's *d* based on the mean and the standard deviation of participants' test scores at the pretest and the two posttests. Cohen's *d* for a repeated-measure design can be defined as:

$$\text{Cohen's } d_{rm} = \frac{\bar{x}_1 - \bar{x}_2}{\frac{\sqrt{s_1^2 + s_2^2 - 2rs_1s_2}}{2(1-r)}}$$

where $\bar{x}_1$and $\bar{x}_2$ are the means of test scores at two time points (here, scores on the pretest and the immediate or the delayed posttest, respectively); s_1^2 and s_2^2 are the corresponding standard deviation for the two means; r is the correlation between tests on the two time points. Unfortunately, no primary study in the dataset reported the correlation between the pretest and the two posttests, which is often argued as a reason why Cohen's *d* for repeated measures hardly becomes the subject of meta-analysis in general (Harrer et al., 2019). However, test-practice effects are by nature a type of within-subjects phenomena, and the use of Cohen's *d* for a repeated-measure design was unavoidable for the purpose of this study. To circumvent the issue, I committed to an assumption that an identical test (or very similar tests) administered to the same sample of participants at two time points would correlate positively (which is true in general) and derivided five Cohen's *d*s for each test contrast of Pretest vs. Immediate and Pretest vs. Delayed in the range of r = .1, .25, .5, .75, and .9. Since there were 100 pairs of test contrasts (see above), this resulted in a total of 500 effect sizes. Statistically synthesizing effect sizes also requires the standard error of the effect size estimate (so the contribution of each effect size can be weighted depending on its accuracy), and standard errors for Cohen's *d* can be approximated using the sample size (Shadish & Haddock, 2009). I hence estimated the standard error for each of the 500 effect sizes as:

$$SE = \frac{n_1 + n_2}{n_1 n_2} + \frac{d^2}{2(n_1 + n_2)}$$

Table 1. The coding scheme of the current study.

Categories	Variable	Levels
Identification	Study ID	1–21
	Group ID	1–23
	Data ID	1–100
	Author(s)	The name of author(s)
	Year of publication	1993–2014
	Publication status	Published, Unpublished
Research design	Sample size	The sample size of the control group
	Number of posttests	The number of posttests conducted
	Test contrast	Pretest vs. Immediate, Pretest vs. Delayed
Effect size	Mean	The mean of participants at a pretest and a posttest
	Standard deviation	The standard deviation on a pretest and a posttest
	Cohen's *d*	The effect size derived from the mean and the standard deviation

Categories	Variable	Levels
	Standard error of Cohen’s *d*	The standard error of Cohen’s *d* derived from the sample size
Outcome measure	Test name	Name of the outcome measure
	Mode	Comprehension, Production
	Degree of awareness	Feel (0), Rule (1)
	Time available	Pressured (0), Unpressured (1)
	Focus of attention	Meaning (0), Form (1)
	Metalinguistic knowledge	No (0), Yes (1)
	Total score	The total score of explicitness based on Awareness, Time pressure, Attention, and Metalinguistic focus
	Outcome type	Explicit (total score = 3–4), Implicit (total score (0–2)

Outcome measures used at the pretest and the posttests were also coded for the mode of language use (comprehension or production) and whether the measures were explicit and implicit tests of L2 knowledge (explicit or implicit measures). I followed Ellis’s (2005) conceptual framework for analyzing explicit and implicit L2 knowledge measures based on four test design features: (a) degree of awareness referred to the extent to which learners become aware of linguistic knowledge being tested (feel or rule); (b) time available referred to whether or not learners are pressured to produce a response (pressured or unpressured); (c) focus of attention referred to whether or not the test draws learner attention to meaning or linguistic form (meaning or form); (d) metalinguistic knowledge referred to whether or not the use of metalanguage was encouraged to complete the test (yes or no). Being aware of the linguistic feature tested (rule), an unpressured test (unpressured), focus on form (form), and the use of metalanguage (yes) was considered to contribute to the explicitness of the outcome measures. One point was awarded for each of these explicitness features, and outcome measures that scored in the range of 0–2 were categorized as implicit measures, whereas those with a total score of 3–4 were coded as explicit measures. Table 2 shows some examples of how each outcome measure was coded. Although recent research suggested that the timed grammaticality judgment test may tap into speeded (or automatized) access to explicit knowledge rather than implicit knowledge (e.g., Suzuki, 2017; Vafaee et al., 2017), I chose to stick to the original categorization of Ellis (2005), as there is no reliable way of coding such fine-grained characterization of L2 knowledge.

Table 2. The categorization of outcome measures based on the four features from Ellis (2005).

Measure	Awareness	Time	Attention	Metalanguage	Total
Elicited imitation test	0 (Feel)	0 (Pressured)	0 (Meaning)	0 (No)	0 (Implicit)
Timed GJT	0 (Feel)	0 (Pressured)	1 (Form)	0 (No)	1 (Implicit)
Sentence combining test	0 (Feel)	1 (Pressured)	1 (Form)	0 (No)	2 (Implicit)
Error correction test	1 (Rule)	1 (Unpressured)	1 (Form)	0 (No)	3 (Explicit)
Untimed GJT	1 (Rule)	1 (Unpressured)	1 (Form)	1 (Yes)	4 (Explicit)

Note. GJT = grammaticality judgment test.

The entire dataset was coded through an iterative process in which the coded features were added, deleted, or redefined as necessary. Since the four designs features from Ellis (2005) were considered of high inference, I recruited a second coder to ascertain the inter-coder reliability on those features. The agreement was near perfect between the two coders, with agreement rates higher than 95% across the four features. The entire coding sheet is available at: https://github.com/maieryo/research/blob/materials/Coding_Maie(2022)SLS.xlsx

Analysis

The analysis consisted of two steps. First, I pooled all effect sizes together and estimated the population estimate of the test-practice effects in L2 grammar instruction research. Since most of the 21 reports contributed more than one effect size, I used a multilevel random-effects model to estimate the population effect size (Cheung, 2014; Harrer et al., 2019). When multiple data points come from the same participant samples or studies, they often correlate with each other, and are said to have a nested data structure, i.e., data points are nested within participant samples or studies. However, this nested data structure cannot be handled by traditional meta-analytic models such as the single-level fixed- and random-effects models (Hedges & Vevea, 1998) because analyzing such data violates the statistical assumption of data independence. Cheung (2014) discusses that when a meta-analyst faces this kind of data dependence, one can either (a) ignore the dependence, (b) take the average of effect sizes, (c) select a single estimate per each study, or (d) directly model the dependence in statistical models. The first three approaches are unjustified because they either violate the assumption of data independence or lose (potentially) useful information that can be contributed by each data point. The use of multilevel models, on the other hand, allowed me to account for the nested structure in the dataset.

As the effect sizes were roughly distributed in a normal distribution (see Figure 2), I constructed a statistical model that assumed that each data point (i.e., Cohen's *d*) came from a normal distribution with the mean of $\theta_{i,j}$ and the standard deviation of $\sigma_{i,j}$:

$$d_{i,j} \sim \mathrm{Normal}(\theta_{i,j}, \sigma_{i,j}).$$

Here, $d_{i,j}$ was the ith data point in jth study, and $\theta_{i,j}$ was considered the true value of $d_{i,j}$after accounting for sampling error, $\sigma_{i,j}$, which was a known variable because the measurement error corresponded to the standard error of each coded effect size. I then modeled these true values of data points, $\theta_{i,j}$, as drawn from a normal distribution centered at $\mu + a_i + a_j$with the standard deviation of τ:

$$\theta_{i,j} \sim \text{Normal}\left(\mu + a_i + a_j, \tau\right).$$

μ is the population estimate of Cohen's *d* and hence the overall size of Test-practice effects in L2 grammar instruction research. a_i and a_j expressed how much each $\theta_{i,j}$ deviated from the true population estimate (μ) due to unobserved error specific to each data point (a_i) and each study (a_j). These parameters are equivalent to the so-called varying (or random) intercepts in regression analysis. The model is considered a multilevel model in that it accounted for how much each data point varied at the study level and at the data level.

In the moderator analysis, I examined to what extent the mode of language use (comprehension or production) and the type of outcome measures (explicit or implicit measures) moderated the overall testing learning effect. In doing so, I created another model that included the two variables as predictors:

$$\theta_{i,j} \sim \text{Normal}\left(\mu + a_i + a_j + b_{\text{Mode}[i]}x_{\text{Mode}[i]} + b_{\text{Type}[i]}x_{\text{Type}[i]}, \tau\right).$$

$b_{\text{Mode}[i]}$ and $b_{\text{Type}[i]}$ expressed how the overall size of testing learning effect, $\mu + a_i + a_j$, changed depending on whether the mode of language use was comprehension or production and whether the outcome measures were explicit or implicit knowledge measures. The two variables were dummy-coded with comprehension and explicit measures as the baseline category. For both models, I used the *R*-package *brms* (version 2.16.3, Büerkner, 2017) that provided an interface to fit Bayesian models using *Stan* (version 2.21.3, Stan Development Team, 2018). *Stan* is a probabilistic programing language for full Bayesian inference and optimization. In Bayesian analysis, prior knowledge in the form of probability distributions is combined with observed data to produce posterior distributions. For the current analysis, I used weakly informative priors for all model parameters to be estimated. The posterior distribution of the model parameters was estimated through Markov chain Monte Carlo simulation from four chains of 10,000 iterations each, with a warmup period of 1,000 iterations and the amount of thinning being two to reduce auto-correlation of the posterior samples. To check whether each chain converged into model parameters with a stationary distribution, we monitored whether the value of $\hat{R}$ associated with each parameter (as a convergence index) was within the range of $1 \leq \hat{R} \leq 1.1$ (Gelman & Rubin, 1992). For both steps of the analysis, I carried out a separate analysis for the two test contrasts (Pretest−Immediate posttest and Pretest−Delayed posttest). Because the effect sizes can vary depending on the five levels of correlation between the pretest and the immediate or the delayed posttest, the number of statistical models increased five folds: 2 (synthesis or moderator analysis) × 2 (two test contrasts) × 5 (five levels of correlation between repeated measures) = 20 models. Although it is more logical to interpret results at each level of the correlation (because it affords a more granular and nuanced view), I will also present the results of the model that is the average of the model at each level of the correlation. This was done through Bayesian model averaging using the *posterior_average* function in the *brms* package (see also Hoetinget al., 1999 for a tutorial). I took the median of the posterior distribution and the highest density interval (i.e., 95% credible interval: CrI) as the parameters' point and interval estimate, respectively.

Results

Preliminary Analysis

Figure 2 shows how the raw effect sizes were distributed at each level of correlation between the pretest and the immediate or the delayed posttest. As discussed above, the effect sizes seemed normally distributed. The distributions in Figure 2 were mostly unimodal and roughly symmetrical on both sides of the mean (shown as the red vertical line in each panel). Although the effect sizes could be both positive and negative, the mean of each panel was consistently positive, suggesting that test-practice effects generally exist in L2 grammar instruction research (but see below for more detailed results). Next, Figure 3 graphically summarizes the relative proportion of levels in each category used in the moderator analysis. Although there was a balanced focus on both comprehension (48%) and production skills (52%), implicit measures occupied a higher proportion of the dataset, which remained true for all sub-level features that went into the categorization of the outcome measures: Degree of awareness (Rule = 35%; Feel = 65%), Time available (Pressured = 57%; Unpressured = 43%), Focus of attention (Meaning = 59%; Form = 41%), and Metalinguistic knowledge (Yes = 30%; No = 70%). The larger representation of implicit measures was a surprising result given the finding in the previous meta-analyses that reported the predominance of explicit measures in L2 grammar instruction research (see Doughty, 2003).

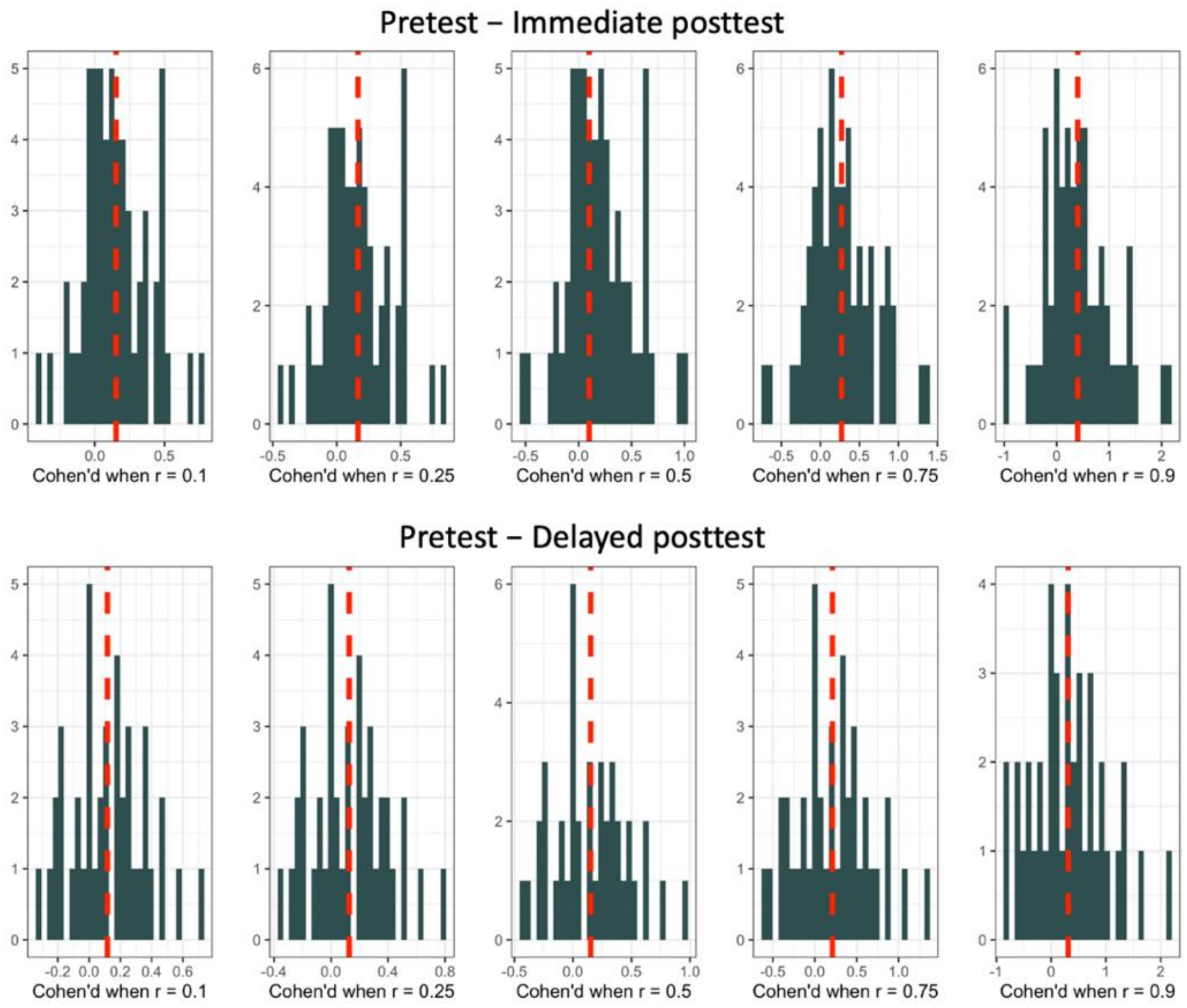

Figure 2. The descriptive distribution of Cohen’s *d*.

Note. The vertical lines indicate the mean of the distribution.

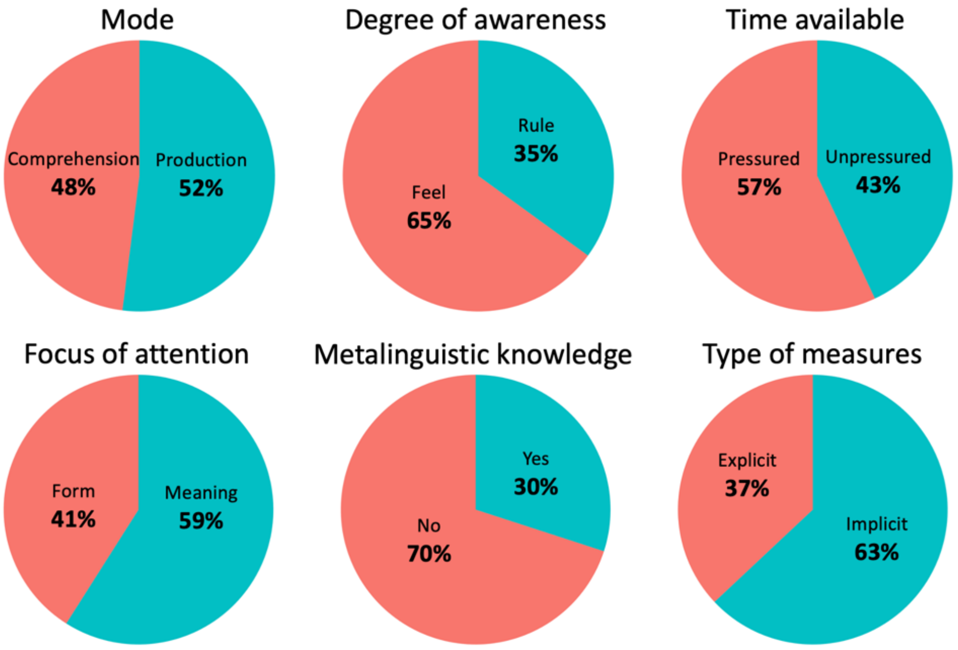

Figure 3. The relative ratio of levels in the coded category.

Overall Synthesis

The population estimate of the synthesized effect sizes is presented in Figure 5 for each level of correlation. The figure also demarcates the magnitude of the effect sizes according to the L2-specific benchmark suggested by Plonsky and Oswald (2014): small (d = 0.2), medium (d = 0.5), and large (d = 0.8). At the immediate posttest, the overall estimates lied around small effect sizes (i.e., d = 0.2), with d = 0.125, 95% CrI [0.031, 0.218] when r = .1; d = 0.136, 95% CrI [0.041, 0.023] when r = .25; d = 0.165, 95% CrI [0.068, 0.263] when r = .5; d = 0.232, 95% CrI [0.114, 0.350] when r = .75, but the magnitude of the test-practice effect became closer to a medium size when the correlation was r = .9 (d = 0.368, 95% CrI [0.196, 0.541]). When the five models (for each level of the correlation) were combined through Bayesian model averaging, the mean population estimate was d = 0.124, 95% CrI [0.031, 0.218], and the posterior probability of the estimate being larger than 0 (i.e., Cohen's *d* > 0) was .995. This suggested that from the pretest to the immediate posttest, control groups in L2 grammar instruction research generally improve by 0.124 points in the standard deviation unit, and with the probability of 99.5%, test-practice effects exist in control groups.

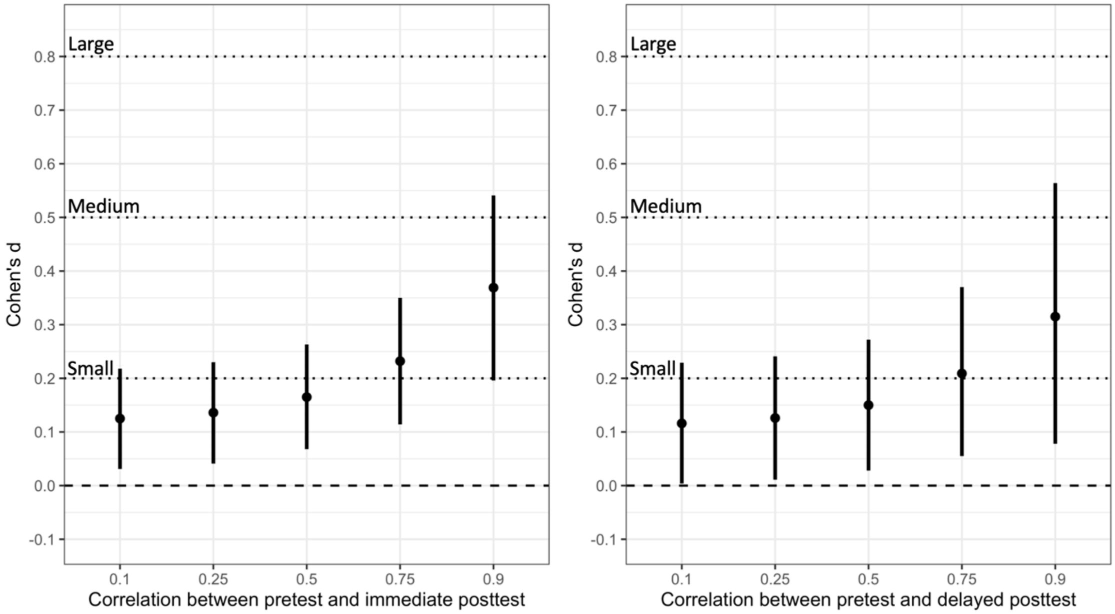

Figure 5. The posterior estimates of the overall test-practice effect

At the delayed posttest, the population estimates were comparable to those of the immediate posttest, but they carried more uncertainty as shown by the wider 95% credible intervals (Figure 5, the right panel), especially when the correlation between the repeated measures became high(er). Despite the caveat of uncertainty, the overall estimates of the effect sizes were largely small in size (*d* = 0.116, 95% CrI [0.004, 0.229] when *r* = .1; *d* = 0.126, 95% CrI [0.011, 0.241] when *r* = .25; *d* = 0.150, 95% CrI [0.028, 0.272] when *r* = .5; *d* = 0.209, 95% CrI [0.055, 0.370] when *r* = .75), but the magnitude of the effects became in between a small and medium effect size when the correlation was *r* = .9 (*d* = 0.315, 95% CrI [0.078, 0.564]). The mean population estimate in the averaged model was *d* = 0.115, 95% CrI [0.031, 0.218], and the posterior probability of the population estimate being larger than 0 was .978. This indicated that from the pretest to the delayed posttest, participants in a control group generally improve by 0.115 points in the standard deviation unit, and Test-practice effects exist with the probability of 97.8%.

To investigate how much variation in the dataset was explained by the parameters that encoded data- (a_i) and study-specific (a_j) error, I calculated the heterogeneity statistic, I^2. I followed the formula by Cheung (2014) for the data- and study-specific heterogeneity:

$$I^2_{\text{data}} = \frac{\tau^2_{\text{data}}}{\tau^2_{\text{data}} + \tau^2_{\text{study}} + \tilde{v}}$$

$$I^2_{\text{study}} = \frac{\tau^2_{\text{study}}}{\tau^2_{\text{data}} + \tau^2_{\text{study}} + \tilde{v}}$$

τ^2_{data} and τ^2_{study} were the variance explained by the data-specific and study-specific random effects (a_i and a_j) and $\tilde{v}$ was the variance of sampling error ($\sigma_{i,j}$). While I^2_{data} described the percentage of variation in the raw effect sizes that was explained by how much the true values deviated from the population mean (μ) due to the data-specific error, I^2_{study} represented the amount of variation that was explained by how the true values deviated from the overall mean due to the study-specific error. The average I^2_{data} across the five models was .549 at the immediate posttest and .517 at the delayed posttest, and the average I^2_{study} was 0.331 and 0.353. This suggested that the data-specific error caused about 51-55% of the variance in the dataset, and 33-35% was due to the study-specific error. This also meant that the models did not explain 84-90% of the variance, but this was understandable as the models only included one fixed effect parameter, i.e., μ, the overall mean. In regression analysis, this is equivalent to modeling data only with an intercept parameter. According to Higgins and Thompson (2002), I^2 (the sum of the two heterogeneity statistics) larger than .75 indicates that substantial heterogeneity is present in data. This was a very good reason to conduct the moderator analysis as some of the heterogeneity might be explained by study-specific differences in the modes of language use (comprehension and production) and the type of outcome measures (explicit or implicit measures)

Moderation Analysis

Results of the moderation analysis are presented in Figure 6 for the mode of language use and Figure 7 for the type of outcome measures as predictor variables.

Comprehension versus Production. At the immediate posttest, the effect sizes were estimated to be larger when the outcome measures drew on comprehension skills (*d* = 0.179, 95% CrI [0.035, 0.320] when *r* = .1; *d* = 0.195, 95% CrI [0.049, 0.335] when *r* = .25; *d* = 0.233, 95% CrI [0.086, 0.379] when *r* = .5; *d* = 0.319, 95% CrI [0.149, 0.490] when *r* = .75; *d* = 0.466, 95% CrI [0.228, 0.718] when *r* = .9) than on production skills (*d* = 0.056, 95% CrI [-0.096, 0.212] when *r* = .1; *d* = 0.061, 95% CrI [-0.098, 0.211] when *r* = .25; *d* = 0.072, 95% CrI [-0.092, 0.229] when *r* = .5; *d* = 0.101, 95% CrI [-0.083, 0.293] when *r* = .75; *d* = 0.183, 95% CrI [-0.088, 0.457] when *r* = .9). Glancing through the estimates, the effect sizes were 2.5-3.2 times larger for comprehension. In absolute terms, the mean difference between the two modes was b_{Mode} = -0.123, 95% CrI [-0.346, 0.103] (see the model formulation in the Analysis section), suggesting that the effect sizes were overall smaller for production skills by 0.123 standard deviation units. Despite the difference, the effect sizes remained in the range of a small to medium effect for both skill types (see Figure 6, the upper panel).

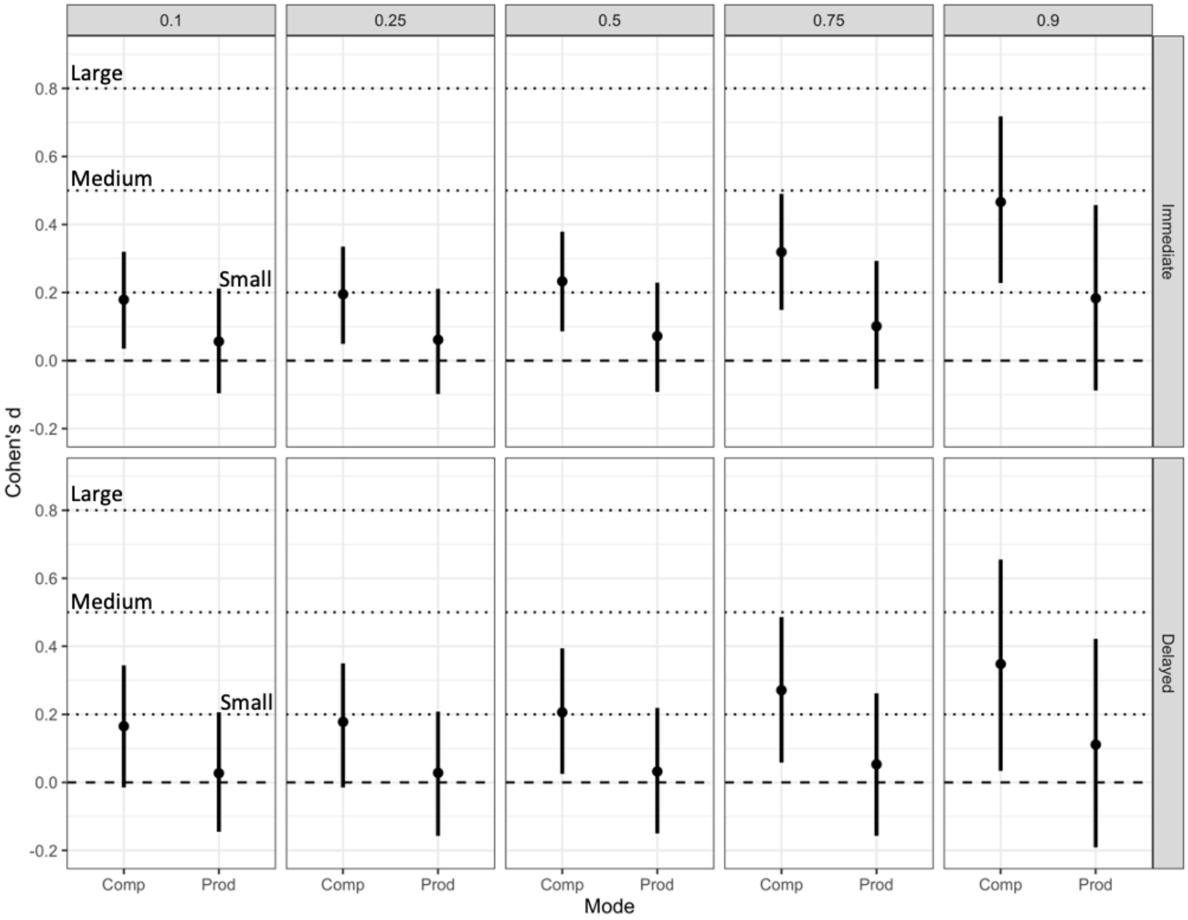

Figure 6. The moderation effect of the mode of language use.

At the delayed posttest, the difference between comprehension and production skills was still evident except that the effects were relatively smaller. While the effect sizes for comprehension skills still lied between a small to medium effect (*d* = 0.165, 95% CrI [-0.015, 0.344] when *r* = .1; *d* = 0.178, 95% CrI [-0.010, 0.350] when *r* = .25; *d* = 0.206, 95% CrI [0.025, 0.394] when *r* = .5; *d* = 0.271, 95% CrI [0.058, 0.486] when *r* = .75; *d* = 0.348, 95% CrI [0.034, 0.655] when *r* = .9), those for production skills were close to zero (*d* = 0.027, 95% CrI [-0.145, 0.206] when *r* = .1; *d* = 0.028, 95% CrI [-0.157, 0.208] when *r* = .25; *d* = 0.032, 95% CrI [-0.150, 0.219] when *r* = .5; *d* = 0.053, 95% CrI [-0.157, 0.262] when *r* = .75; *d* = 0.111, 95% CrI [-0.191, 0.433] when *r* = .9), suggesting that unless scores on the pretest and the delayed posttest are highly correlated, test-practice effects may be negligible (at least) for production skills. In the averaged model, the difference between comprehension and production was b_{Mode} = -0.136, 95% CrI [-0.401, 0.136], which indicated that that the test-practice effects were overall weaker for production skills by 0.136 standard deviations.

Explicit versus Implicit. At the immediate posttest, the synthesized effect sizes (across the five levels of levels) were consistently positive for explicit measures of L2 knowledge, but their size was nearly zero, suggesting that testing learning effects may be negligible for this type of outcome measures: *d* = 0.061, 95% CrI [-0.107, 0.224] when *r* = .1; *d* = 0.066, 95% CrI [-0.109, 0.228] when *r* = .25; *d* = 0.076, 95% CrI [-0.094, 0.255] when *r* = .5; *d* = 0.095, 95% CrI [-0.103, 0.299] when *r* = .75; *d* = 0.138, 95% CrI [-0.154, 0.426] when *r* = .9. On the other hand, there were small effects for implicit measures (*d* = 0.173, 95% CrI [0.042, 0.312] when *r* = .1; *d* = 0.188, 95% CrI [0.058, 0.312] when *r* = .25; *d* = 0.229, 95% CrI [0.091, 0.366] when *r* = .5; *d* = 0.324, 95% CrI [0.163, 0.481] when *r* = .75) and when the correlation between the pretest and the delayed posttest was *r* = .9, the effect increased to a medium size (*d* = 0.512, 95% CrI [0.285, 0.735]). Overall, the estimates suggested that test-practice effects were 2.3-3.4 times larger for implicit measures. In absolute terms, the averaged difference between explicit and implicit measures was b_{Type} = 0.110, 95% CrI [-0.118, 0.349], which meant that the effects were overall stronger for implicit measures by 0.110 standard deviations.

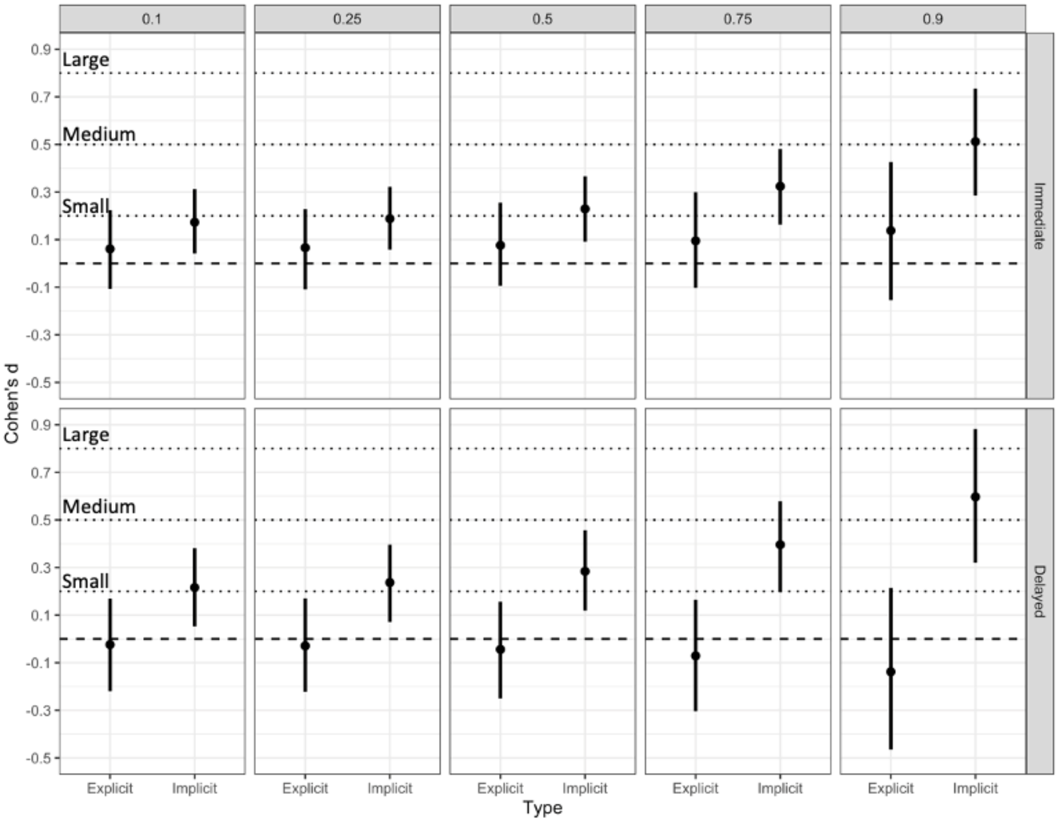

Figure 7. The moderation effect of the type of outcome measures.

At the delayed posttest, the difference between the two outcome measures seemed more noticeable possibly because the effect sizes became smaller for explicit measures. In fact, the effect sizes for this type of measures were negative at all five levels of correlation (*d* = -0.024, 95% CrI [-0.220, 0.169] when *r* = .1; *d* = -0.029, 95% CrI [-0.222, 0.170] when *r* = .25; *d* = -0.044, 95% CrI [-0.250, 0.156] when *r* = .5; *d* = -0.071, 95% CrI [-0.304, 0.165] when *r* = .75; *d* = -0.138, 95% CrI [-0.465, 0.214] when *r* = .9). However, their size was very close zero, suggesting that testing learning effects were almost non-existent for explicit measures (rather than negative). For implicit measures, however, the test-practice effects surely existed especially when the correlation between the pretest and the delayed posttest was high (*d* = 0.216, 95% CrI

[0.053, 0.381] when r = .1; d = 0.237, 95% CrI [0.071, 0.396] when r = .25; d = 0.284, 95% CrI [0.119, 0.457] when r = .5; d = 0.396, 95% CrI [0.196, 0.579] when r = .75; d = 0.597, 95% CrI [0.320, 0.882] when r = .9). The difference between explicit and implicit measures of L2 knowledge was b_{Type} = 0.240, 95% CrI [-0.029, 0.517]. This meant that the effect sizes were more severe for implicit measures by 0.240 points in the standard deviation unit.

Discussions and Conclusions

In this pilot study, I examined the overall test-practice effect in L2 grammar instruction research and investigated whether two moderating variables, the mode of language use (comprehension or production) and the type of outcome measures (explicit or implicit measures), changed the size of the test-practice effect. Figure 8 and 9 present a comparative summary of the synthesized effect sizes at the immediate and the delayed posttest, respectively. Overall, test-practice effects in L2 grammar instruction research are small in size (d = 0.124 and 0.115 at the immediate and the delayed posttest, respectively), although their effects exist with great certainty (99.5% and 97.8% at the immediate and the delayed posttest, respectively). However, when one uses outcome measures that draw on production skills or explicit knowledge of L2, the effect sizes were estimated to be near zero, which meant that test-practice effects on these measures can be negligible. On the other hand, outcome measures of comprehension skills or implicit knowledge are more sensitive to test-practice effects, showing small to medium effect sizes. In a meta-analysis by Goo et al. (2015), the overall effectiveness of L2 instruction was estimated as 1.095 at the immediate posttest and 0.841 at the delayed posttest, but they did not take into account the potential test-practice effects.[3] Considering the overall size of the test-practice effect found in the current study, these estimates can be re-interpreted by subtracting the overall size of test-practice effects we found in this study from the effects of L2 instruction: d = 0.971 (i.e., 1.095–0.124) and 0.726 (0.841–0.115). Similarly, one can return to previous studies of L2 grammar instruction research (that did not recruit a control group) and re-interpret their findings.

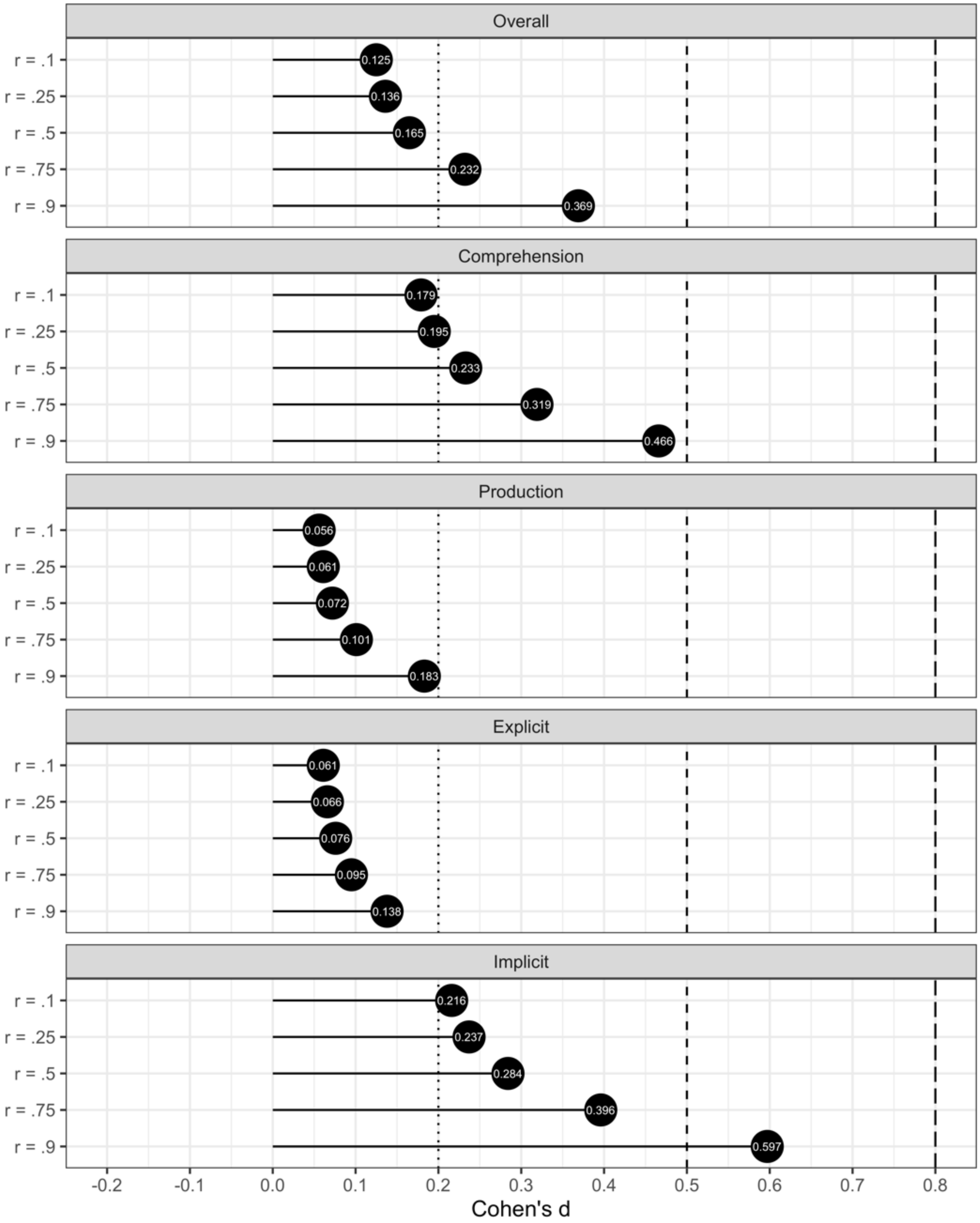

Figure 8. The comparative summary of the effect sizes at the immediate posttest.

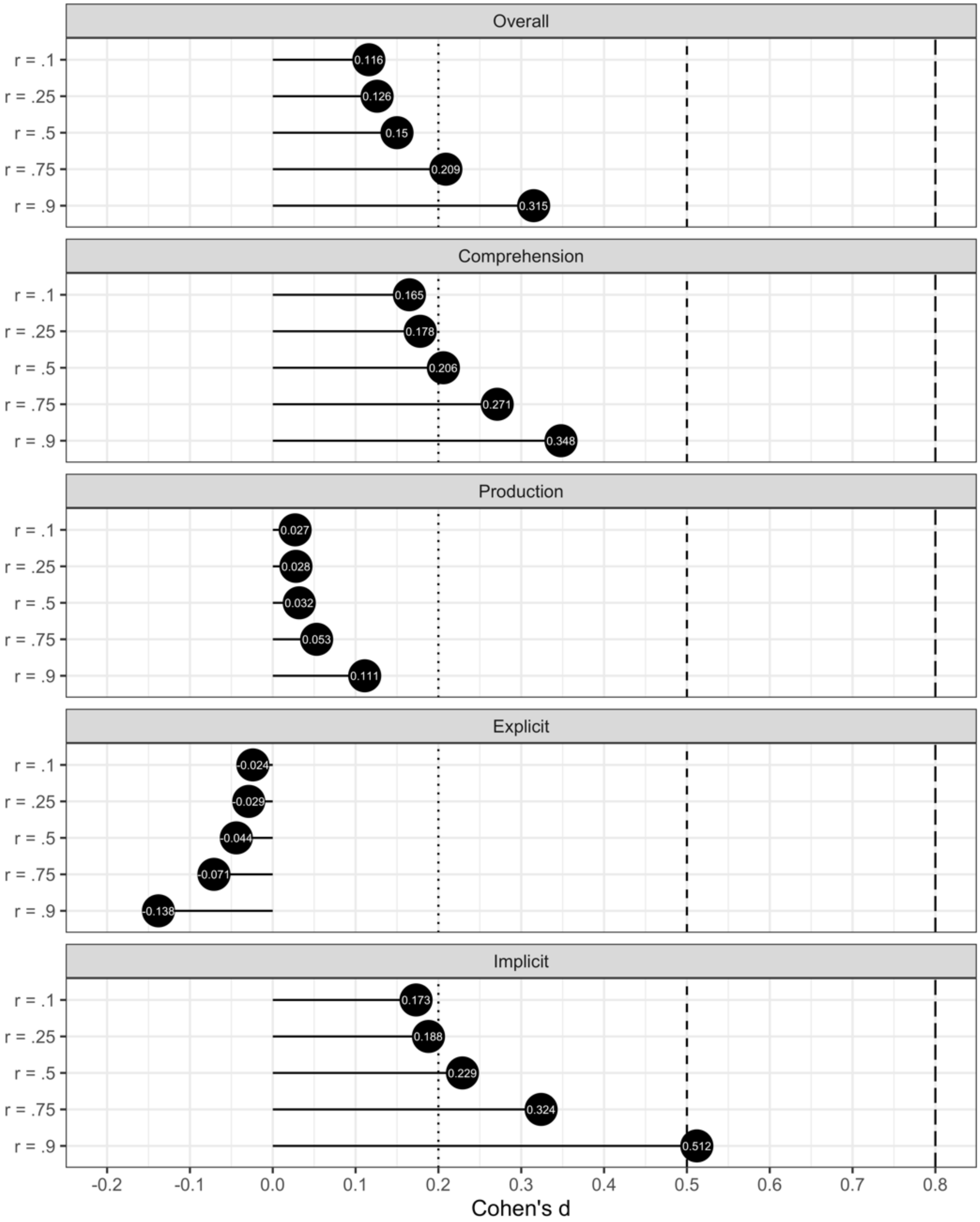

Figure 9. The comparative summary of the effect sizes at the delayed posttest.

It is interesting to speculate why the synthesized effects were larger when the mode of language use involved comprehension rather than production and the type of outcome measures was implicit rather than explicit. When it comes to the first comparison, comprehension measures may be more amenable to test-practice effects because acquiring comprehension skills requires less effort than acquiring production skills. Producing a grammatical structure is a (more) attentionally demanding process (Kormos, 2006) unless learners automatize the process, which was certainly not true for participants in the primary studies included in the current meta-analysis because the participants did not receive any instruction. Regarding the second comparison, implicit knowledge measures may be more amenable to test-practice effects because most implicit measures (e.g., the elicited imitation test and the oral/written production test) gauged performance using (whatever) knowledge of the target grammatical structure, whereas explicit knowledge measures (e.g., the grammaticality judgment test and the metalinguistic knowledge test) mostly assessed learners' knowledge of the target structure. According to skill acquisition theory, it is implicit (or procedural) knowledge that improves with practice rather than explicit (or declarative) knowledge (DeKeyser, 2020).

In this pilot study, I considered only two variables. However, many other variables can potentially influence the size of the test-practice effect. These new variables are expected to be addressed in the main project. The first obvious candidate is the number of test items included in each outcome measure, as this variable can serve as a direct proxy for the number of learning opportunities L2 learners experience besides instructional treatments (see Suga, 2022). It is not hard to imagine that the higher the number of test items, the larger the test-practice effect becomes. However, does the test-practice effect linearly increase as a function of the number of test items; or does it require a certain number of items to show an effect? Furthermore, the number of test items should also interact with learners' prior knowledge, especially explicit and declarative knowledge of the target language features. As discussed in Suga (2022), if test items serve as additional opportunities to practice the target feature, having prior knowledge likely facilitates the test-practice effect because such learning condition exactly matches the one advocated by skill acquisition theory: "declarative knowledge is used as a clutch during practice to attain procedural knowledge" (Suzuki, Nakata, & DeKeyser, 2019, p. 553).

The second candidate is types of grammar structures being tested. L2 research has shown that different kinds of grammar structures are learned (more or less) differently in terms of the rate of learning and the cognitive processes involved (see DeKeyser, 2005 for a review). If this is the case, the same phenomena should appear for the test-practice effect. In this respect, Suga (2022) investigated whether morphological and syntactic structures (in English) differentially react to the test-practice effect, but the same question can be addressed more robustly through a meta-analysis. While the current pilot study was restricted in terms of the scope of synthesis (i.e., the test-only control group) and the number of moderator variables included (i.e., the mode of language use and the type of outcome measures), the intention of the main project is to carry out a meta-analysis of test-practice effects in both control and comparison groups including a wider array of moderator variables in the analysis. I believe such a study is critical as it not only reveals whether Test-practice effects exist in L2 grammar instruction research but also helps us more validly interpret the effectiveness of L2 instruction shown in previous studies.

References

Cheung, M. W.-L. (2014). Modeling dependent effect sizes with three-level meta-analyses: A structural equation modeling approach. *Psychological Methods*, *19*(2), 211–229. https://doi.org/10.1037/a0032968

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

DeKeyser, R. M. (2005). What makes learning second-language grammar difficult? A review of issues. *Language Learning*, *55*(s1), 1–25. https://doi.org/10.1111/j.0023-8333.2005.00294.x

DeKeyser, R. M. (2020). Skill acquisition theory. In B. VanPatten, G. D. Keating, & S. Wulff (Eds.), *Theories in second language acquisition. An introduction* (3rd ed., pp. 83–104). New York, NY: Routledge.

Doughty, C. J. (2003). Instructed SLA: Constraints, compensation, and enhancement. In C. J. Doughty & M. H. Long (Eds.), *The handbook of second language acquisition* (pp. 256–310). Malden, MA: Wiley-Blackwell.

Ellis, R., Loewen, S., & Erlam, R. (2006). Implicit and explicit corrective feedback and the acquisition of L2 grammar. *Studies in Second Language Acquisition*, *28*(2), 339–368. https://doi.org/10.1017/S0272263106060141

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis* (3rd ed.). Boca Raton, FL: CRC Press.

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, *7*, 457–511. https://doi.org/10.1214/ss/1177011136

Goo, J., Granena, G., Yilmaz, Y., & Novella, M. (2015). Implicit and explicit instruction in L2 learning: Norris & Ortega (2000) revisited and updated. In P. Rebuschat (Ed.), *Implicit and explicit learning of languages* (pp. 443–482). https://doi.org/10.1075/sibil.48.18goo

Harrer, M., Cuijpers, P., Furukawa, T. A, & Ebert, D. D. (2019). *Doing meta-analysis in R: A hands on guide*.

https://bookdown.org/MathiasHarrer/Doing_Meta_Analysis_in_R/

Higgins, J. P., & Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine*, *21*(11), 1539–1558. https://doi.org/10.1002/sim.1186

Hoeting, J. A., Madigan, D., Raftery, A. E., Volinsky, C. T. (1999). Bayesian model averaging: A tutorial. *Statistical Science*, *14*(4), 382–417. https://doi.org/10.1214/ss/1009212519

Kang, E. Y., & Han, Z. (2015). The efficacy of written corrective feedback in improving L2 written accuracy: A meta-analysis. *The Modern Language Journal*, *99*(1), 1–18. https://doi.org/10.1111/modl.12189

Kao, C.-W. (2013). Effects of focused feedback on the acquisition of two English articles. *TESL-EJ*, *17*(1), 1–15.

Kormos, J. (2006). *Speech production and second language acquisition*. Mahwah, NJ: Lawrence Erlbaum Associates.

Li, S. (2010). The effectiveness of oral corrective feedback in SLA: A meta-analysis. *Language Learning*, *60*(2), 309–365. https://doi.org/10.1111/j.1467-9922.2010.00561.x

Li, S., Ellis, R., & Zhu, Y. (2016). Task-based versus task-supported language instruction: An experimental study. *Annual Review of Applied Linguistics*, *36*, 205–229. https://doi.org/10.1017/S0267190515000069

Lim, C. S., & Renandya, W. A. (2020). Efficacy of written corrective feedback in written instruction: A meta-analysis. *TESL-EJ*, *24*(3), 1–26.

Loewen, S. (2020). *Introduction to instructed second language acquisition* (2nd ed.). London, UK: Routledge.

Loewen, S., Erlam, R., & Ellis, R. (2009). The incidental acquisition of third person *-s* as implicit and explicit knowledge. In R. Ellis, S. Loewen, C. Elder, R. Erlam, J. Philp, & H. Reinders (Eds.), *Implicit and explicit knowledge in second language learning, testing and teaching,* (pp. 262-280). Bristol, UK: Multilingual Matters.

Long, M. H. (1983). Does second language instruction make a difference? A review of research. *TESOL Quarterly*, *17*(3), 359–382. https://doi.org/10.2307/3586253

Long, M. H. (1988). Instructed interlanguage development. In L. M. Beebe (Ed.), *Second language acquisition: Multiple perspectives* (pp. 115–141). Cambridge, MA: Newbury House.

Lyster, R., & Saito, K. (2010). Oral corrective feedback in classroom SLA: A meta-analysis. *Studies in Second Language Acquisition*, *32*(s2), 265–302. https://doi.org/10.1017/S0272263109990520

Norris, J. M., & Ortega, L. (2000). Effectiveness of L2 instruction: A research synthesis and quantitative meta-analysis. *Language Learning*, *50*(3), 417–528. https://doi.org/10.1111/0023-8333.00136

Plonsky, L., & Oswald, F. L. (2014). How big is “big”? Interpreting effect sizes in L2 research. *Language Learning*, *64*(4), 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, NY: Routledge.

Shadish, W. R., & Haddock, C. K. (2009). Combining estimates of effect size. In L. V. Hedges, H. Copper, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (pp. 257−277). New York, NY: Russell Sage.

Sok, S., Kang, E. Y., Han, Z. (2019). Thirty-five years of ISLA on form-focused instruction: A methodological synthesis. *Language Teaching Research*, *23*(4), 403−427. https://doi.org/10.1177%2F1362168818776673

Spada, N., & Tomita, Y. (2010). Interactions between type of instruction and type of language feature: A meta-analysis. *Language Learning*, *60*(2), 263−308. https://doi.org/10.1111/J.1467-9922.2010.00562.X

Stan Development Team (2018). Stan: A C++ library for programming and sampling. Retrieved from http://mc-stan.org

Suga, K. (2022, March). Potential Test-practice effects of an oral elicited imitation test: Methodological considerations for form-focused instruction studies. Paper presented at the American Association for Applied Linguistics, Pittsburg, PA.

Suzuki, Y. (2017). Validity of new measures of implicit knowledge: Distinguishing implicit knowledge from automatized explicit knowledge. *Applied Psycholinguistics*, *38*(5), 1229−1261. https://doi.org/10.1017/S014271641700011X

Suzuki, Y., Nakata, T., & DeKeyser, R. (2019). Optimizing second language practice in the classroom: Perspectives from cognitive psychology. *The Modern Language Journal*, *103*, 551−561. https://doi.org/10.1111/modl.12582

Vafaee, P., Suzuki, Y., & Kachinske (2017). Validating grammaticality judgment tests: Evidence from two new psycholinguistic measures. *Studies in Second Language Acquisition*, *39*(1), 59−95. https://doi.org/10.1017/S0272263115000455

Footnote

1. Here, I'm assuming that the experimental group and the control group are comparable concerning confounding variables, such as general proficiency in the target language, prior knowledge of target language structures, and cognitive, affective, and conative individual difference variables that can affect how well learners learn L2 in general and through instruction. This problem of confounding variables can be preemptively avoided by randomly assigning participants to the experimental and control groups. However, random assignments are extremely difficult when one recruits intact classes rather than individual participants (and the classes themselves are assigned to the groups). Sok, Kang, and Han (2019) reported that in thirty-five years of FFI research, only 54% used the random assignment. The assumption that the experimental and the control group are comparable a priori may not be as valid as most L2 researchers believe. Alternatively, one can measure those confounding variables and include them in the analysis to see if individual differences in the confounding variables moderate the effect of instruction.
2. An anonymous reviewer pointed out that there are in fact many labels for the phenomenon, including practice effects, test-retest effects, and test-learning effects. While the use of the term test-learning effects implicates that one assumes that an effect from multiple testing constitutes actual learning of the content (i.e., L2 development), the terms such as practice effects and test-retest effects simply mean that participants become more skilled in taking the test. In this article, I am using the term test-practice effects to refer to general improvements from a pretest to posttests and do not argue whether or not such improvement constitutes actual learning of the content (but cf. Suga, 2022).

3. Goo et al. (2015) reported the effect size in the form of Hedge's *g*, which is very similar to Cohen's *d* except that Hedge's *g* is slightly more robust against biases in estimates due to small sample sizes. Both statistics are on the standard deviation unit and hence are directly comparable.