*Studies in Second Language Acquisition* (2021), 1–26
doi:10.1017/S0272263121000516

# *Methods Forum*

## *P*-CURVING AS A SAFEGUARD AGAINST *P*-HACKING IN SLA RESEARCH

### A CASE STUDY

***Seth Lindstromberg*** *

*Hilderstone College*

**Abstract**

It is important to be able to identify research results likely to have been arrived at by means of "*p*-hacking," a common term for research and reporting practices (such as the selective reporting of results) that are biased toward finding $p < \alpha$. This paper discusses and demonstrates "*p*-curving," a means of checking a set of primary studies within a specific research stream for signs of *p*-hacking. A salient feature of *p*-curving is that it is based entirely on significant *p*-values. Because of the potential usefulness of *p*-curving and because it has been little used by SLA researchers, a case study illustrates the construction and analysis of a *p*-curve as a complement to meta-analysis. The focal *p*-curve in this study relates to published (quasi)experimental studies that addressed the research hypothesis that for low and middle proficiency learners L1 glosses facilitate vocabulary learning during reading better than L2 glosses do.

## INTRODUCTION

Threats posed to the integrity and usefulness of research by publication bias are staple topics of discussions of statistical meta-analysis (e.g., Borenstein, et al., 2009; Carter et al., 2019; Hedges, 1992; Rothstein et al., 2005) and of the apparently widespread difficulty of replicating statistically significant findings (Fidler & Wilcox, 2018). In such discussions the term *publication bias* has chiefly been used to refer to the tendency of publication gatekeepers (especially editors and reviewers) to approve submitted reports of statistically significant findings in preference to equally good reports of inconclusive

I am grateful to the lead authors of Kim et al. (2020) and Yanagisawa et al. (2020) for answering questions and to the lead author of Kim et al. for supplying two papers that I could find no other way of obtaining. At various stages of its development this paper benefited immensely from comments, suggestions, and corrections from Frank Boers, Tessa Woodward, three anonymous reviewers, and an editor.

* Correspondence concerning this article should be addressed to Seth Lindstromberg, Hilderstone College, St Peters Road, Broadstairs, Kent, CT10 2JW, United Kingdom. Email: lindstromberg@gmail.com; sethl@hilderstone.ac.uk

findings. Discussions of publication bias frequently include discussion both of "file drawering" and a large category of questionable practices referred to as "*p*-hacking" (Head et al., 2015; Simonsohn et al., 2014a, 2014b, 2015, 2019), "hidden flexibility" (Chambers, 2017), or "selective reporting" (Head et al., 2015; Simonsohn et al., 2014a), among several other terms. File-drawering and *p*-hacking are both most likely to occur when researchers assume that publication bias exists and believe that their work has been or may be affected by it.

*File drawering* occurs when a researcher either makes no attempt to publish or else gives up attempting to publish a study or part of a study that yielded $p > \alpha$ for a result in the expected direction or that yielded a result of $p > \alpha$ or $p < \alpha$ in the *un*expected direction. Because file-drawered studies tend to be ones that found relatively small effects, a consequence of file drawering is that findings of small effects remain disproportionately unavailable to other researchers and to meta-analysts. A result of this is that published estimates of effect sizes tend to be inflated, with potential adverse effects on theory and research agendas (e.g., Borenstein et al., 2009; Hedges, 1992; Rosenthal, 1979; Rothstein et al., 2005).

The terms *p-hacking*, *selective reporting*, and so on denote types of biased selection among available options in data collection, data analysis, and reporting of results—or, more precisely, the "misreporting of true effect sizes in published studies [that] occurs when researchers try out several statistical analyses and/or data eligibility specifications and then selectively report those that produce significant results" (Head et al., 2015, p. 1). Because of its especially frequent use in the recent literature, the term used most often in this article is *p-hacking*. It is likely, though, that much *p*-hacking occurs not because researchers mean to cheat but because they hold mistaken beliefs about what analytic procedures are permissible (Gelman & Loken, 2013).

Wicherts et al. (2016) discussed 34 types of *p*-hacking, for example: (a) dealing with outliers in an ad hoc manner or, worse yet, dealing them only after finding $p > \alpha$ (hereafter $\alpha = .05$), then retesting the data in the hope of finding $p < .05$, and, finally, reporting only the most favorable result (Bakker & Wicherts, 2014; Pollet & van der Meij, 2017); (b) failing to report tested dependent measures or covariates for which $p > .05$; and (c) carrying out two or more different tests of statistical significance on the same data (e.g., a *t*-test and a Wilcoxon–Mann–Whitney test [WMW]) and, if only one test yields $p < .05$, then reporting only that *p*-value.

## CONTEXT, GRADATIONS, AND INCIDENCE OF *P*-HACKING

Taking as background the evidently common situation in which researchers feel pressured for a study to yield at least one significant result, Gelman and Loken (2013) described how a researcher engaged in a quantitative study must choose among options that arise at many separate stages, from conception to reporting. Within this state of affairs, vividly characterized by Gelman and Loken as "a garden of forking paths," it is often easy for a researcher to make a sequence of decisions—each decision being defensible in itself but also arbitrary and motivated—that leads eventually to a statistically significant result in some part of the experimental design.[1] Indeed, the scope for hidden flexibility in decision making is typically so wide that, as demonstrated by Simmons et al. (2011), researchers who engage in determined *p*-hacking in multiple seemingly tiny ways can succeed in

finding $p < .05$ for almost any research hypothesis whatsoever. It is plausible that such a result is especially easy to achieve in a field such as SLA where research hypotheses typically have the potential to involve a multitude of variables of intersecting types (e.g., pedagogical, temporal, cultural, item-related, and learner-related) (cf., Roettger, 2019).

While it has been speculated that *p*-hacking in applied research tends to be modest in degree (Simonsohn et al., 2015), results of a survey carried out by John et al. (2012) indicate that *p*-hacking has been knowingly engaged in from time to time by an appreciable minority of researchers in psychology, an important feeder field of SLA. A particularly striking result of the survey of John and colleagues is that a sizeable proportion of the researchers who admitted to having engaged in types of *p*-hacking seemed to be unaware of the extent to which those practices are types of scientific misconduct that can degrade the quality of shared knowledge within their research community. To give an example of a different sort, Bakker and Wicherts (2011) examined 281 randomly selected articles in high- and low-impact journals of psychology and found that about 18% of the reported *p*-values were inconsistent with the reported degrees of freedom and/or the test statistic. Bakker and Wicherts found additionally that erroneous *p*-values tended to be in a direction favorable to researchers' hypotheses. One possibility then is that researchers, without necessarily being aware of it, are selective in how carefully they proofread their results sections.

## THE CHALLENGE OF DISTINGUISHING SIGNS OF *P*-HACKING FROM SIGNS OF FILE-DRAWERING

There are at least a dozen methods for estimating and perhaps also correcting for the incidence of file-drawering in a body of related quantitative studies (e.g., Carter et al., 2019; van Aert & van Assen, 2021).[2] Among the best known of these methods are "fail-safe *N*," for estimating the number of effect sizes in the file-drawer (Rosenthal, 1979); "funnel plotting," for detecting so-called small study effects (Egger et al., 1997; Light & Pillemer, 1984); "Egger's regression test," for testing funnel plot symmetry (Egger et al., 1997); and "trim and fill," for correcting funnel plot asymmetry (Borenstein et al., 2009; Duval & Tweedie, 2000). In contrast, methods devised specifically to detect *p*-hacking have been lacking. For one thing, procedures conceived for the purpose of detecting file-drawering and/or compensating for it are unable to distinguish signs of *p*-hacking from signs of file-drawering. This is true, for instance, of methods commonly used by SLA meta-analysts. One example is trim and fill combined with funnel plotting.[3] Essentially, researchers using existing methods for detecting and mitigating impacts of file-drawering must assume that there was no *p*-hacking in the primary studies being investigated (e.g., Carter et al., 2019). Yet there is abundant evidence that *p*-hacking is a threat that must be taken seriously (e.g., Bakker & Wicherts, 2011; Simmons et al., 2011). In short, there is a need for a separate method that detects *p*-hacking. Fortunately, there *is* such a method. It is hoped that this article will accelerate its use by researchers of SLA.

## INTRODUCING *P*-CURVING

The present study introduces a comparatively simple and effective method—dubbed "*p*-curving" or "*p*-curve analysis" by Simonsohn et al. (2014a)—that was developed

specifically to detect signs of *p*-hacking. A salient feature of this method is that it is based entirely on a distribution of *significant p*-values. The prime rationale for this restriction is that a published statistically significant finding cannot have been affected by file drawering because being affected by file drawering means not being published. Consequently, any bias detectable in a set of significant *p*-values is likely to stem from *p*-hacking. A second rationale for the focus on significant *p*-values is that, for any given research hypothesis, it is comparatively easy to find a reasonably complete sample of studies that yielded a relevant significant *p*-value. In contrast, owing to file drawering, it is often impossible to be certain of having found even a majority of relevant studies where $p > .05$ (Hedges, 1992; Simonsohn et al., 2014a).

## *P*-CURVES

### *DEFINITIONS*

The term *p*-curve denotes a "distribution of statistically significant *p*-values for a set of independent findings" (Simonsohn et al., 2014a, p. 535), where "independent findings" are findings made in separate studies based on responses from different study participants. However, for extra clarity this article sometimes replaces the term *p-curve* with the longer expression *0 to .05 p-curve* so that the expression *0 to 1 p-curve* can be used to refer to a distribution of significant or nonsignificant *p*-values for a set of independent findings.

*P*-curves are important because inspection of a *p*-curve can reveal signs of *p*-hacking. This is possible because it is known what *p*-curves should look like in the absence of *p*-hacking. But what should any *p*-curve look like in any circumstance? The next section begins the process of answering that question.

### *A* P-*CURVE WHEN THE NULL HYPOTHESIS OF NO EFFECT IS TRUE AND THERE WAS NO* P-*HACKING*

When a null hypothesis (*H0*) is true and a relevant test of statistical significance is unbiased, a vastly long sequence of properly conducted and properly analyzed studies independently addressing that *H0* can normally be expected to yield a distribution of *p*-values that is either completely "uniform" (i.e., flat) or very nearly so (e.g., Head et al., 2015; Lakens, 2014; Simonsohn et al., 2014a).[4] In such a distribution, 5% of the *p*-values lie between 0 and .05; a further 5% lie between .05 and .10; and so on right up to and including the interval between .95 and 1, as indicated by Figure 1. (R script for all simulations and figures is given in the final section of the online Supplementary Materials and on the Open Science Framework at osf.io/vsqg8)

Because the *p*-values shown in Figure 1 were derived by performing independent samples (IS) *t*-tests on samples drawn randomly from a single population in which the true effect size is 0, it was a matter of chance whether the mean of a sample X turned out to be greater than the mean of a sample Y or vice versa. The *size* of any particular mean difference was a matter of chance as well. Thus, every significant *p*-value represented in Figure 1 is significant solely because of random sampling variation. Hence, one useful interpretation of a flat 0 to .05 *p*-curve is that a study represented in that interval would fail

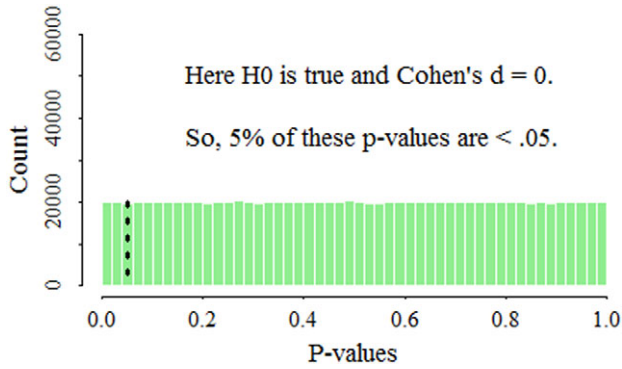


FIGURE 1. The distribution of *p*-values from Student's independent samples *t*-test applied to one million pairs of samples, where each sample "X" and each sample "Y" ($n_X = n_Y = 20$) was a random draw from a normal distribution having $Mn = 10$ and $SD = 1$. *Note*: The dashed line marks $p = .05$.

to yield a significant *p*-value 19 times out of 20 if it was directly replicated with new learners.

### *A P-CURVE WHEN THE H0 OF NO EFFECT IS FALSE AND THERE WAS NO P-HACKING*

If the *H0* is *false* (e.g., if $True\ Mean_X \neq True\ Mean_Y$), then the distribution of *p*-values will show some degree of right-skew. In other words, most of the *p*-values will be heaped up near 0 with a thin tail of larger *p*-values extending to the right (Figure 2): Technically, the *p*-values will have an "exponential" distribution. When sample sizes are large (e.g., of size 200 rather than 20) and when the *H0* is conspicuously false (e.g., when two population means are very different), then the right skew of a *p*-curve may reach the point of being essentially L-shaped, or even nearly columnar (e.g., Lakens, 2014). Importantly, this rightward skew of *p*-values will also be evident within the range 0 to .05 (Figure 3).

### *A P-CURVE WHEN THE H0 IS TRUE AND THERE WAS P-HACKING*

Having considered the shape of *p*-curves when there was no *p*-hacking, we are ready to consider what *p*-curves may look like when *p*-hacking did occur. In the first of the following two simulated scenarios the built-in true effect was 0; however, concurrently two types of *p*-hacking were at play. One type was ad hoc exclusion of outliers. The other was an egregious version of the practice of applying two significance tests to the same data in the hope of increasing the chance of finding $p < .05$. Specifically, the simulation mimicked a scenario in which a researcher conducts a *t*-test on two independent samples using software that automatically supplements *t*-test results with results of the WMW test. In the scenario, the researcher chooses the *t*-test *p*-value if $p < .05$. But if that *p* value is > .05, the researcher chooses the *p*-value from the WMW test if it is < .05. But if neither test has yielded $p < .05$, the simulated researcher deletes the minimum and maximum values from each dataset and then tests the difference between the two trimmed datasets,

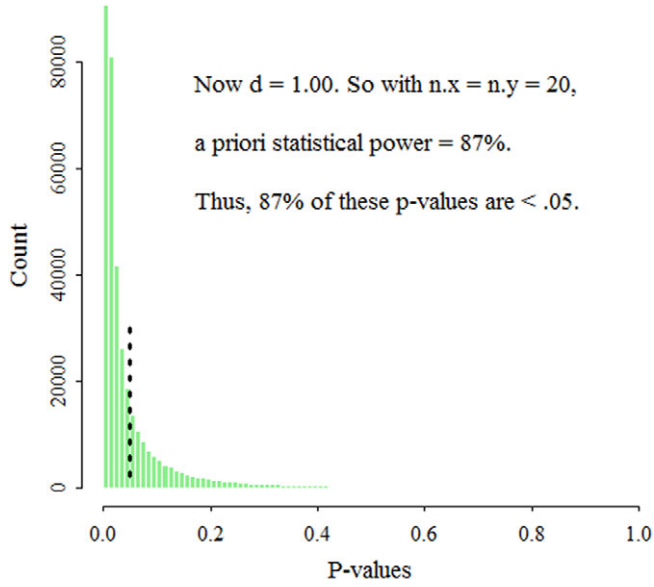


FIGURE 2. One million *p*-values generated in the same way as for Figure 1 except that $Mn_X = 10$ and $Mn_Y = 11$. *Note*: The dashed line marks $p = .05$.

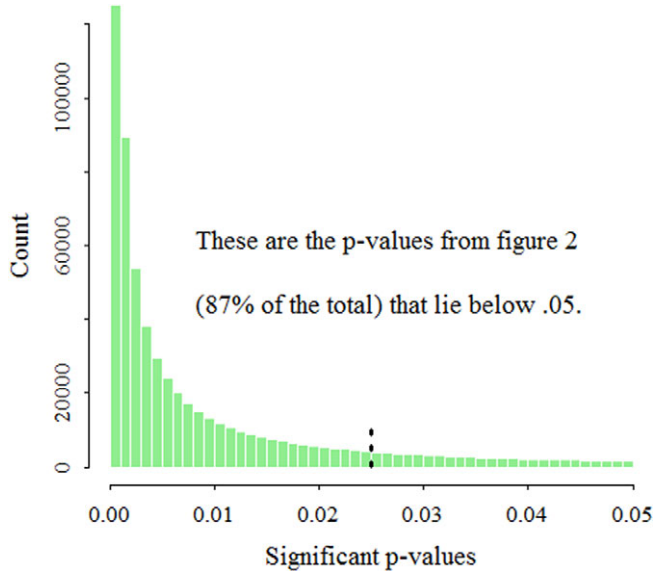


FIGURE 3. The 0 to .05 portion of the *p* distribution depicted in Figure 2. The exponential distribution is still conspicuous. *Note*: The dotted line marks $p = .025$.

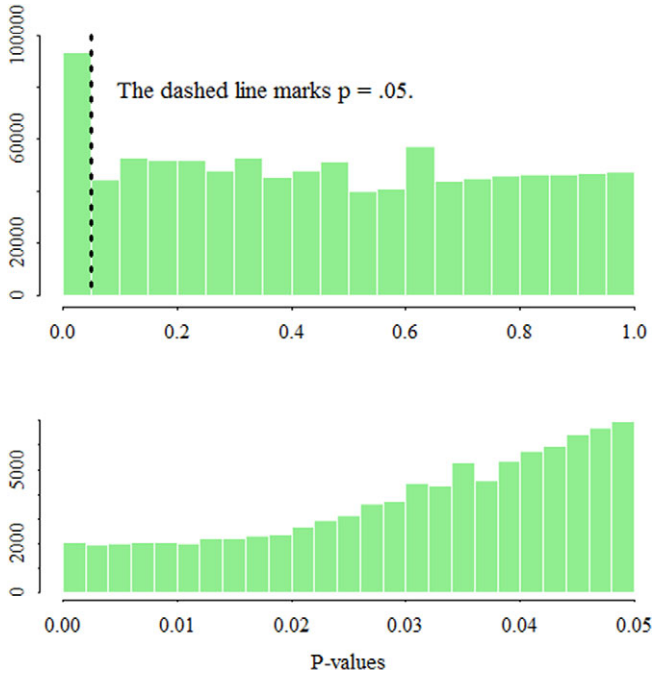


FIGURE 4. The combined effect of two types of *p*-hacking when $d = 0$.

choosing the *p*-value from the *t*-test if $p < .05$. But if that *p*-value is > .05, the researcher chooses the *p*-value from the WMW test whatever it happens to be. Further, the simulated researcher reports the chosen significance test without mentioning any of the others. The scenario was simulated one million times with the following settings: $n_X = n_Y = 25$; $Mn_X = Mn_Y = 20$; $SD_X = 4.50$; $SD_Y = 4.00$.

The *p*-curve resulting from the new simulation is depicted in Figure 4. In the upper histogram in that figure it can be seen that on average there are almost twice as many *p*-values in the 0↔.05 column, or "bin," as in any of the other 19 bins. The excess *p*-values in the 0↔.05 bin represent type I errors (i.e., false findings). However, it is the lower histogram in Figure 4 that displays the clearest sign of *p*-hacking—namely, the distribution of statistically significant *p*-values is *left* skewed.

### *A P-CURVE WHEN THE H0 IS FALSE AND THERE WAS P-HACKING*

An additional simulation was carried out to illustrate *p*-hacking when there is a true effect that is large enough to be pedagogically significant. In this simulation the settings for the X and Y scores were as for the previous simulation except that now $Mn_X = 20$ and $Mn_Y = 22$. Thus, $d_{population} = 0.47$. In both histograms in Figure 5 the existence of an appreciable effect is signalled by the disproportionately tall bins of *p*-values near 0. In the *lower* histogram, concurrent *p*-hacking is indicated by the fairly consistent left skew between about .02 and .05.

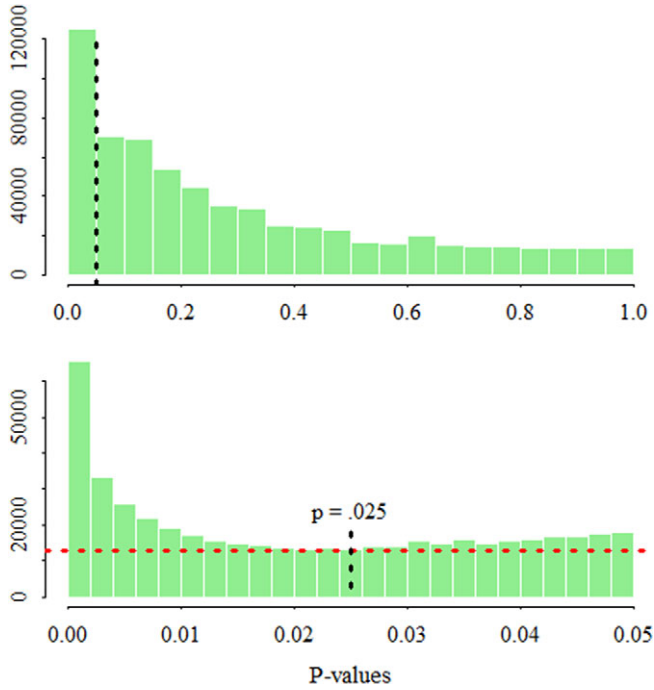


FIGURE 5. This display illustrates the combined impact of two types of *p*-hacking along with a true effect of moderate size. *Notes*: In the upper histogram the dotted line marks $p = .05$. The horizontal line in the lower histogram serves to underscore the concavity of the 0 to .05 *p*-curve.

Although a region of left skew in a 0 to .05 *p*-curve (as in the lower histogram in Figure 5) is a classic indicator of *p*-hacking, it is important to remain aware that the shape of a 0 to .05 *p*-curve is determined by multiple factors (Bishop & Thomas, 2016; Gelman, 2018; Hartgerink, 2017; Lakens, 2014). Among these factors, the size of the true effect is likely to be especially important. Factors that tend to be of lesser importance include the type(s) of statistical significance test used in the underlying primary studies (e.g., Welch's IS *t*-test vs. Student's IS *t*-test) and the type of raw data in those studies (e.g., continuous vs. discrete). Also important are the sizes of the samples in the underlying primary studies and the stability of the true effect from context to context. The comparative importance of *p*-hacking depends on the incidence of *p*-hacking and the type(s) of *p*-hacking involved. Also, each of the *p*-curves illustrated in this article so far was based on one million *p*-values. Consequently, the outlines of these *p*-curves would remain essentially stable if the simulations were repeated with different sequences of random numbers. However, in practical research *p*-curves will be based on far fewer *p*-values, meaning that the contour of any given *p*-curve may be strongly influenced by random sampling variation. The point being made here is that interpretation of real world *p*-curves calls for caution.

It is relevant that left skew can result not only from *p*-hacking but also from an apparently small number of inadvisable practices that have nothing to do with trying to change $p > .05$ into $p < .05$. Erdfelder and Heck (2019) considered a scenario in which a

researcher uses ANCOVA and finds $p < .05$. But the researcher goes on to use ANOVA or a *t*-test also, because compared to ANCOVA these procedures are better known, they require fewer assumptions, and they can be described in comparatively few words. Finally, if the ANOVA (or the *t*-test) yields $p < .05$ then, for simplicity and brevity, the researcher reports this *p*-value without mentioning the use of ANCOVA. If this scenario occurred in an appreciable proportion of the primary studies underlying a *p*-curve, that curve might show some left skew simply because a significant *p*-value from an ANOVA or a *t*-test will often be larger than a corresponding ANCOVA *p*-value. It may be doubted though that this scenario is at all common in SLA research.

## USES OF *P*-CURVING WITH RESPECT TO THE POSSIBLE PRESENCE OF *P*-HACKING

Although *p*-curving can have a second use that is touched on in a later section, this article focuses on its use to gain an impression of the overall trustworthiness and replicability of a set of published statistically significant results. There are several types of situation in which SLA quantitative researchers should consider *p*-curving for the purpose just mentioned, given that existence of *p*-hacking is a realistic possibility.

First, *p*-curving can enhance the planning of a research program. For example, Lakens (2018) demonstrated how *p*-curving can enable a researcher to distinguish between a research path that is promising and one that is not. In Lakens's study, each path had to do with a behavioral effect of priming, but the types of priming were not the same. Both types had been experimentally studied multiple times, and for each type statistically significant positive results had been published. For each of these two sets of published results Lakens constructed a 0 to .05 *p*-curve. He found that one of the *p*-curves was strongly indicative of *p*-hacking whereas the other showed the hallmark of solid evidential value—namely, conspicuous right skew. Lakens's point was that *p*-curving can make the difference between embarking on program of successful research and embarking on a research program that leads a succession of results that are inconclusive and unpublishable.

A second use of *p*-curving is exemplified by a very large-scale, multisite study carried out by Dick et al. (2019) to address the hypothesis that childhood SLA is beneficial to development of (cognitive) executive function. Past results regarding this hypothesis had been mixed. Dick et al. subjected the relevant statistically significant *p*-values to a *p*-curve analysis and found little sign of a substantively significant true effect when possible effects of *p*-hacking were taken into account. Dick et al. used this result as one component of their tripartite basis for stating that their study revealed no persuasive evidence of a bilingual advantage for executive function in children aged from 9 to 10.

Simonsohn et al. (2014a, p. 535) suggested a range of additional uses of *p*-curving, as follows. First, *p*-curving "will be useful to anyone who finds it useful to know whether a given set of significant findings is merely the result of selective reporting." Second, *p*-curving may help researchers to "decide which articles or literatures to read attentively, or as a way to assess which set of contradictory findings is more likely to be correct." Third, a *p*-curve analysis may suggest an explanation of inconsistencies between one's own findings and findings published previously. Fourth, "reviewers may use *p*-curve [*sic*] to decide whether to ask authors to attempt a direct replication prior to accepting a

manuscript for publication." Fifth, reviewers and editors may conduct *p*-curve analyses to gain an impression of the reliability of findings relating to a research hypothesis addressed in a submitted research report. Simonsohn et al. suggested additionally that *p*-curving might be applied to a single article that reports many *p*-values. However, discussion of this last use is beyond the scope of this article. Finally, *p*-curving is a tool for assessing the evidential value of statistically significant findings in a stream of research covered by one or more published quantitative meta-analyses or narrative literature reviews. A later section of this article presents a case study of *p*-curving with respect to two published quantitative meta-analyses.

## THE ASSUMPTION OF HOMOGENEITY IN *P*-CURVING

Developers of the *p*-curve method for detecting *p*-hacking have stressed that it works best when an effect of interest is relatively homogeneous at population level; and they have pointed out that problems of interpretation arise if the magnitude of an effect varies notably from context to context (Simonsohn et al., 2014a). However, in applied research the degree to which an effect is homogeneous at population level is generally unknowable. Accordingly, a prospective *p*-curver will need to make an informed judgement about the homogeneity (vs. heterogeneity) of the effect of interest before undertaking a *p*-curve analysis. That said, many useful statistical procedures depend on the ultimately unverifiable tenability of assumptions about population characteristics: ANOVA is a prime example.

The homogeneity or heterogeneity of studied effects has been a major concern in recent studies of *p*-curving for the purpose of estimating true effect sizes (e.g., McShane et al., 2016). The consensus seems to be that *p*-curve methods for *this* purpose must be adapted so that they take account of both significant and nonsignificant *p*-values because otherwise the resulting estimates are likely to be inaccurate and probably inflated whenever the true effects are heterogeneous (Carter et al., 2019; McShane et al., 2016; van Aert & van Assen, 2021).[5] It seems that no such adjustment is necessary or even feasible with respect to the *p*-curve method for detecting *p*-hacking.

## PREPARATORY STEPS OF *P*-CURVING USING THE ONLINE *P*-CURVE APP

First steps of *p*-curving include choosing a research hypothesis and identifying a set of studies relating to that hypothesis. Then it is necessary to find the studies that yielded a directly relevant significant *p*-value—typically just one *p*-value per study. (An important exception is specified in the following text.) Note that in *p*-curving there is no averaging of related *p*-values analogous to the standard meta-analytic practice of averaging related effect sizes. Note too that if one's working hypothesis happens to be that a true effect is positive, it is necessary to set aside any significant *p*-values that correspond to a negative effect. The opposite would be done if a negative effect has been hypothesized. Finally, it can be difficult and perhaps impossible to interpret a *p*-curve that includes *p*-values relating to different research hypotheses (e.g., Simonsohn et al., 2014a).

It is possible to create an informative *p*-curve histogram by using reported *p*-values. However, reported *p*-values are likely to be rounded to two decimal places, and some may be reported as, say, "< .05" or "< .001." *P*-values that are desirably exact can usually be

TABLE 1. The basic steps of *p*-curving

| |
|---|
| Settle on a research hypothesis—no more than one hypothesis per *p*-curve. |
| ⇩ |
| Find relevant studies that produced a significant *p*-value. Verify that the *p*-value is statistically significant by recalculating it from reported descriptive statistics. |
| ⇩ |
| For each *p*-value it is necessary to have the degrees of freedom and the test statistic (e.g., *t*). |
| ⇩ |
| Put these two statistics in the right format, for example: *t*(88)=2.1145. No spaces, please. Do not round down statistics to the usual small number of decimal places.* |
| ⇩ |
| Arrange the statistics in a single-spaced, comma-free column—with one set of statistics per row as in Table 3. |
| ⇩ |
| Copy the column and enter it into the *p*-curve app. |
| ⇩ |
| Click on "Make the *p*-curve." |
| ⇩ |
| Study the graphical display and the printed output. Most of the key print output is either embedded in the graphical display or is given just below it. |
| ⇩ |
| Interpret what you see. Are there any red flags? If so, how serious do they look? |

*Further examples of correctly formatted sets of statistics are as follows: *F*(1,100)=.2460, *f*(2,2102)=4.4573, *Z*=3.45706, chi2(1)=9.1255

recalculated from reported descriptive statistics. Also, recalculation of *p*-values occasionally enables a reporting error to be corrected. However, there is a free-to-use "*p*-curve app" that produces superior print and graphical output and is supported by copious explanatory material (Simonsohn et al., 2015, 2017) (http://p-curve.com/). Use of this app does not require recalculation of *p*-values because these are calculated by the app. Throughout the rest of this article use of this app will be assumed. The essential steps of *p*-curving with the *p*-curve app are summarized in Table 1.

Because ANOVA has been so heavily used by SLA researchers (Lindstromberg, 2016; Plonsky, 2013; Plonsky & Gass, 2011), it is important for *p*-curvers to bear it in mind that an *attenuating interaction* (in which an effect merely varies between being strong and weak) must be handled differently than a more radical *reversing* (or crossover) interaction. In the case of an attenuating interaction, the only *p*-value that is used is the one for the interaction. If an interaction is reversing, the *p*-curver uses the *p*-values for the simple effects (NB: not the main effects).

For an attenuating interaction the rationale is as follows:

> When a researcher investigates the attenuation of an effect, the interaction term will tend to have a larger *p* value than the simple effect (because the latter is tested against a null of 0 with no noise, and the former is tested against another parameter estimated with noise). Because publication [bias] incentivizes the interaction to be significant, *p* values for the simple effect will usually have to be much smaller than .05 in order for the study to be published. This means that, for the simple effect, even *p* values below .05 will be censored [i.e., absent] from the published record and, thus, not uniformly distributed under the null [hypothesis]. (Simonsohn et al., 2014a, p. 543, note 16)[6]

Finally, Simonsohn et al. (2014a, p. 544) stated that 10 *p*-values can be the basis of an informative *p*-curve analysis, given adequate statistical power. However, their statistical power simulations assumed only 20 participants per group. (Coincidentally, 20 learners per group has been typical of between-subjects designs in SLA experimental research [Lindstromberg, 2016; Plonsky, 2013].) Recall, though, that for any given effect size, statistical power will rise with the number of participants per group, all else being equal. It is therefore plausible that somewhat fewer than 10 studies may suffice if the average group size in the studies is considerably larger than 20.

## GUIDELINES FOR UPHOLDING THE INTEGRITY OF *P*-CURVE ANALYSES

Various sets of guidelines have been proposed for maximizing the integrity of *p*-curve analyses (e.g., BITSS, 2017; Head et al., 2017; Simonsohn et al., 2014b, p. 1151). Two guidelines are particularly prominent in this literature. First, decide in advance which studies to include. Disclose the decision. Adhere to it. Second, be open about choices made in cases of uncertainty. For example, if it is unclear whether a study should be included in an analysis, report results for when the study is included and for when it is not. Fulfillment of these and other published guidelines are illustrated in and in connection with the case study that is reported in the text that follows.

## *P*-CURVING IN SLA AND BILINGUALISM RESEARCH SO FAR

To date *p*-curving has been little used or referred to either by SLA researchers or by researchers of bilingualism. In bilingualism research a conspicuous exception is the study by Dick et al. (2019) that was summarized in a previous section. Also with respect to bilingualism research, Plonsky et al. (2021) referred to use of *p*-curving for the purpose of estimating effect sizes in a meta-analytic context. As to SLA research, Vitta and Al-Hoorie (2020) used the *p*-curve app (Simonsohn et al., 2017) to carry out a *p*-curve analysis as one strand of their standard random effects meta-analysis of empirical studies of the "flipped classroom." Discussion of that *p*-curve analysis is postponed until after presentation of the case study illustrating an application of *p*-curve analysis.

## A *P*-CURVING CASE STUDY: L2 VOCABULARY GLOSSING

### *INTRODUCTION*

Fixed effects and random effects meta-analyses are unable detect *p*-hacking; and, as mentioned, the same is true of ancillary methods such as trim and fill (e.g., Carter et al., 2019; Simonsohn et al., 2014b; van Assen et al., 2015; van Aert & van Assen, 2021). Accordingly, it has been recommended that journals should publish *p*–curve reevaluations of the extent of *p*-hacking in sets of studies covered by published meta-analyses that included no effective checks for signs of *p*-hacking (van Aert et al., 2019). The case study reported in the following text is intended to illustrate this use of *p*-curving in an SLA context. The case study covers primary studies that addressed the possibility that L1 and L2 glosses differ in the degree to which they promote acquisition of word

meanings by learners of elementary or intermediate proficiency. In all these studies learners were asked to demonstrate knowledge of what meaning attaches to a given L2 orthographic form in an immediate or near immediate posttest and, usually, in a delayed posttest as well. Although all the relevant primary studies used two-sided significance tests ($\alpha = .05$), it seems likely that researchers generally expected to find that L1 glossing is comparatively effective. The case study proceeded from the assumption that the researchers did have this expectation.

The decision to build a case study around the previously mentioned research hypothesis was influenced by the following facts. First, vocabulary acquisition is an essential component of SLA. Second, two recent, relevant meta-analyses were available: Kim et al. (2020) and Yanagisawa et al. (2020). Third, these two meta-analyses covered enough acceptable primary studies for *p*-curving to be feasible and informative. Fourth, casual inspection of a large proportion of the primary studies cited by Kim et al. (2020) indicated that all necessary descriptive statistics were likely to be findable.

Importantly, all the candidate primary studies implemented one or more measures to reduce the chance that target words would be known to learners prior to the experimental treatment. These measure were staging a pilot study to vet candidate target words, administering a pretest either to screen out familiar candidate words or to screen out overly proficient learners, and using English-like pseudowords as targets. Thus, even for the studies that had not used random assignment of learners to treatment groups, it was reasonable to follow the original researchers in analyzing only the posttest scores—considering that the chief purpose of this case study is to demonstrate an application of *p*-curving rather than to illustrate ideal experimental design.

### *CONSTRAINTS THAT THE TWO META-ANALYSES IMPOSED ON THE SELECTION OF PRIMARY STUDIES*

Prior to finalizing the list of to-be-included primary studies, it was decided that the *p*-curve analysis would include no studies other than the ones covered by Kim et al. (2020) and Yanagisawa et al. (2020). It was envisaged that this inclusion rule could serve as a transparent way to keep the case study from growing much larger than is necessary to illustrate an application of *p*-curving. It should be emphasized, though, that in the majority of practical applications it would be preferable for a *p*-curve analysis to include all relevant published studies findable anywhere.

This case study's coverage of primary studies was indirectly limited by the decision of Kim et al. (2020) and Yanagisawa et al (2020) to exclude studies that were not reported in English and that did not use English as the target L2. Of the various other limitations on coverage which this case study inherited from Kim et al. and Yanagisawa et al., only one is questionable. Namely, Kim et al. excluded quasi-experimental studies that had not included use of a test of statistical significance test (generally one-way ANOVA) to confirm that treatment groups were equivalent prior to the treatment. A potentially serious drawback of this policy is that it excludes all primary studies carried out by researchers who know that this sort of baseline equivalence testing is ineffective and potentially misleading (Lindstromberg, 2016; Norris, 2015). Valid equivalence testing works very differently (e.g., Lakens et al., 2018).

### *CHOICES DETERMINING STUDY SELECTION*

In line with previously mentioned purposes and limitations of *p*-curving the following choices were made:

First, the case study would include only primary studies having an experimental design yielding score sets that can in principle be compared by using a two-sided IS *t*-test. This was partly because of the comparative scarcity of relevant primary studies having a within-subjects design (Yanagisawa et al., 2020)—that is, a design in which all subjects experience all conditions—and partly because it is not straightforward to equate results stemming from such a design with results from a between-subjects design (e.g., Plonsky & Oswald, 2014). In particular, a *p*-value from a specific application of the paired *t*-test can relate to several different versions of *d*, each with its own value (Westfall, 2016). Second, the acceptance of a study was conditional on the availability of the relevant sample sizes, means, and SDs. Third, each study must have included at least one group of learners described as being mostly at elementary, low-intermediate, or intermediate level. This limitation was due to the unresolved matter of whether the positive effect of L1 glossing is appreciably attenuated in the case of highly proficient L2 learners.[7] Two exclusions were made solely because some or all test scores were described as having come from postintermediate learners. Fourth, because the relevant goal of *p*-curving is detection of *p*-hacking and because *p*-hacking is presumed to be a consequence of publication bias, it followed that the case study would only include studies reported in a named print or online journal.

Table A of the Supplementary Materials provides an overview of the 10 primary studies that survived the selection process plus an additional study (Kazerouni & Rassaei, 2016) that was not included in the meta-analyses of Kim et al. (2020) and Yanagisawa et al. (2020). Although apparently acceptable overall, Kazerouni and Rassaei's study included neither random assignment of learners to groups nor BE testing. For these reasons that study would not have been eligible for inclusion in the meta-analysis of Kim et al. Kazerouni and Rassaei's study is touched on again further in the text that follows. Finally, Table B in the Supplementary Materials lists candidate primary studies that were excluded from the present case study and gives reasons for their exclusion.

### *SELECTION OF COMPARISONS WITHIN PRIMARY STUDIES*

In many of the primary studies the variable gloss language (L1 vs. L2) was crossed with another binary variable: gloss type (single word vs. multiple choice), learning mode (intentional vs. incidental), proficiency (low vs. high), or $^{+/-}$picture.

As there seems to be no persuasive evidence that gloss type appreciably moderates the effect of gloss language (Kim et al., 2020; Yanagisawa et al., 2020), the test scores for single word and multiple choice glossing were combined for each gloss language. For much the same reason test scores for the intentional and the incidental learning conditions were also combined for each gloss language. It seems relevant that two of the primary studies directly addressed the relative effects of intentional and incidental learning (see Table 2) but found $p > .05$. Further, $^{+/-}$intentional learning is often operationalized as $^{+/-}$forewarning-of-a-test, yet it cannot be taken for granted that most learners will assume that lack of forewarning is a sure sign that no test will follow (e.g., Barcroft, 2015; Boers,

in press). Finally, while six studies definitely did not include forewarning, for two studies it was not made clear whether there was forewarning or not.

A different policy was adopted regarding proficiency and $^{+/-}$picture, on the assumption that it is comparatively risky to assume that these variables do not appreciably moderate the effect of gloss language. Specifically, when a primary study explicitly distinguished between high- and low-proficiency learners, the case study used only the test scores from the low-proficiency group; and when a study had $^{+}$picture and $^{-}$picture conditions (i.e., Yoshii, 2006), the case study used only the scores for the $^{-}$picture condition. Lastly, with respect to proficiency, Kim et al. (2020) applied the term *beginner* to the learners in a large fraction of the primary studies figuring in the present case study. By my reading of the corresponding original articles, the labels *elementary*, *postbeginner*, or *preintermediate* seem more apt, depending on the study.

### *THE POLICY REGARDING INTERACTIONS*

All the candidate primary studies had a between-groups design at each time of testing. However, most of the studies included a delayed posttest, meaning that there could be within-subjects testing of retention (or forgetting). Some of the studies detected an attenuating interaction involving time. Specifically, less forgetting was associated with L2 glossing than with L1 glossing despite overall superiority of L1 glossing. For these studies the relevant *p*-value was the *p*-value for the interaction between time and gloss language (see Table 2). Further, Ko (2017) observed an attenuating interaction between proficiency and gloss language at the immediate posttest. The interaction *p*-value was used for that study also.

### *CALCULATING THE NECESSARY STATISTICS FOR P-CURVING*

The function *tsum.test* in the R package BSDA (Arnholt, 2017) was used to expedite calculation of *t*-statistics, degrees of freedom, and *p*-values. This function applies Welch's *t*-test, which uses degrees of freedom that may run to several decimal places. Once computed, the key statistics for each study were formatted as described in Table 1 and as shown in the far right column of Table 3.

### *RESULTS OF THE MAIN P-CURVE ANALYSIS AND WHAT THEY MEAN*

A key item in the copious output of the *p*-curve app of Simonsohn et al. (2017) is a 0 to .05 *p*-curve graph to be discussed in a later paragraph. Related to that graph are six tests of statistical significance able to furnish evidence that a *p*-curve reflects a practically significant level of evidential value over and above any distortion by *p*-hacking. The app presents the six significance tests in groups of three. The first group of tests are one-sided tests of the *H0* that the *p*-curve has a flat or left-skewed distribution. For these tests $p < .05$ indicates right skew, which in turn indicates appreciable overall evidential value.

The results from the first three tests are given in Table 4. The binomial test addresses the *H0* that there are as many significant *p*-values above .025 as there are below .025. The directional alternative hypothesis is that most of the *p*-values lie *below* .025. Because this test gave $p < .05$, it is warranted to conclude that the *p*-curve (Figure 6) shows evidential

TABLE 2. Experimental designs of the *p*-curved primary studies

| *Study* | *Design* | *Interaction(s)* |
|---|---|---|
| Arpaci (2016) Δ | Glossing (L1, L2, None) by Time (PT, DPT)[a] | An attenuating interaction was observed, $p < .05$. Forgetting was less for L2 glossing than for L1 glossing. |
| Ertürk (2016) | Glossing (L1, L2, None). There was a one-way ANOVA for each time (PT & DPT). | No interaction was tested. |
| Farvardin & Biria (2012) Δ | Glossing (L1 single, L2 single, L2 MC)[b] by Time (PT, DPT). | An attenuating interaction was observed, $p < .05$: Forgetting was less for L2 glossing than for L1 glossing. |
| Ko (1995) | Learning (+ vs. –Intentional) by Glossing (L1, L2, None). Separate analyses for the two times (PT & DPT). | $P > .05$ for a tested interaction. |
| Ko (2017) Δ | Proficiency (High vs. Low) by Glossing (L1, L2, None, L1+L2). | An unspecified interaction of proficiency & gloss language was tested separately for each time (PT, DPT). At PT the observed interaction was attenuating ($p < .05$): Scores for $\text{Proficiency}^{\text{High}}$ were always superior to $\text{Proficiency}^{\text{Low}}$; but the superiority differed according to gloss language. |
| Mitarai & Aizawa (1999) | Gloss language (L1, L2) by Gloss type (Single, MC) by Time (PT, DPT). | $P > .05$ for interactions. |
| Öztürk & Yorgancı (2017) | Gloss language (L1, L2) by Time (PT, DPT). Paired *t*-tests were used to test retention from PT to DPT; $p < .05$ only for L1. | No interaction was reported. |
| Pishghadam & Ghahari (2011) | Glossing (L1 single, L2 single, L1 MC, L2 MC, None). Separate analyses for PT & DPT. | No interaction was reported. |
| Shiki (2008) | Glossing (L1 single, L2 single, L1 MC, L2 MC). One-way ANOVA. | No interaction was reported. |
| Yoshii (2006) Δ | Gloss language (L1, L2) by Picture (Yes, No) by Time (PT, DPT). | There was an attenuating interaction, $p < .05$: Forgetting was less for L2 glossing than for L1 glossing. |
| Kazerouni & Rassaei (2016)[c] | Learning (+ vs. –Intentional, Control). Gloss language (L1, L2). A separate 1-way ANOVAs for PT and DPT. | No interaction was reported. |

[a]PT = (near) immediate posttest; DPT = delayed posttest.
[b]MC = multiple choice glossing.
[c]This extra study is discussed in a later section of this article.

TABLE 3. The p-curve analysis: Studies and statistics*

| | L1 Glossing | | | L2 Glossing | | | |
|---|---|---|---|---|---|---|---|
| *Study* | *N* | *M* | *SD* | *N* | *M* | *SD* | *Input to the app* |
| Arpaci (2016) | 28 | 12.79 | 1.66 | 28 | 10.07 | 3.42 | F(2,75)=13.09 |
| Ertürk (2016) | 42 | 22.90 | 4.16 | 42 | 18.12 | 6.19 | t(71.794)=4.156 |
| Farvardin & Biria (2012) | 40 | 14.45 | 2.86 | 40 | 11.06 | 2.95 | F(2,96)=9.10 |
| Ko (1995) | 64 | 22.16 | 3.60 | 62 | 18.74 | 5.08 | t(109.66)=4.3479 |
| Ko (2017) | 40 | 4.57 | 2.53 | 42 | 3.17 | 2.54 | F(3,321)=4.35 |
| Mitarai & Aizawa (1999) | 102 | 52.71 | 9.95 | 83 | 46.68 | 7.79 | t(182.74)=4.6203 |
| Öztürk & Yorgancı (2017) | 22 | 17.81 | 3.26 | 22 | 15.00 | 4.33 | t(39.018)=2.4317 |
| Pishghadam & Ghahari (2011) | 84 | 7.51 | 1.70 | 87 | 6.40 | 1.33 | t(157.14)=4.7508 |
| Shiki (2008) | 69 | 3.83 | 1.79 | 74 | 2.53 | 1.81 | t(140.45)=4.3245 |
| Yoshii (2006) | 97 | 2.49 | 2.96 | 98 | 2.22 | 1.86 | F(1,193)=9.50 |

* Some authors gave descriptive statistics to more decimal places than are shown here: Recalculation of degrees of freedom and *t*-statistics was based on the descriptive statistics as authors reported them.

TABLE 4. Results of three statistical significance tests of the null hypothesis is that there is no right skew in the *p*-curve shown in Figure 6

| | |
|---|---|
| *Binomial test comparing the number of significant p-values below .025 with the number of significant p-values above .025* | |
| | $p = .001$ |
| *Stouffer's combination test* | |
| The full *p*-curve test: | $Z = -8.28$, p<.0001 |
| The half *p*-curve test: | $Z = -7.19$, p<.0001 |

value. However, the binomial test is based on values dichotomized as either 1 (for < .025) or as 0 (for > .025). Consequently, this test may lack statistical power to detect *moderate* right skew. Tests two and three in Table 3 are applications of Stouffer's combination test. This test is based on *p*-values that have been transformed but not dichotomized. An application of Stouffer's test can therefore be expected to have more power than a corresponding application of the binomial test. A sufficiently large *Z* score from Stouffer's test signifies a high proportion of small *p*-values (i.e., right skew) in the range at issue. The first Stouffer's test focuses on the 0 to .025 half *p*-curve. Any left skew between 0 and .025 would suggest very drastic *p*-hacking. The second Stouffer's test focuses on the whole 0 to .05 *p*-curve. This test is intended to be comparatively sensitive to moderate *p*-hacking. Table 4 shows that each Stouffer's test rejects its *H0*.

The results given in Table 4 indicate that the 10 focal primary studies have overall evidential value with respect to the research hypothesis addressed by this case study. But approximately how much evidential value? This question brings us to Figure 6, which shows the graph produced by the *p*-curve app. In discussing Figure 6 let us begin with the *dotted* horizontal line near the bottom of the graph. This line depicts the *p*-curve that would be expected if L1 and L2 glossing have the same effect on L2 vocabulary learning. In Figure 6 it can be seen that the observed *p*-curve, represented by the solid line, deviates from the dotted line dramatically. The label "90%" by the left end of the solid line

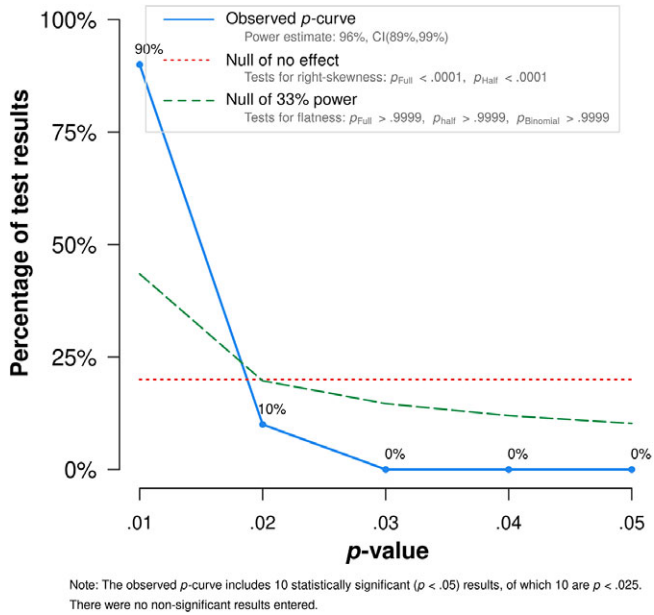


FIGURE 6. *P*-curve produced by the online app of Simonsohn et al. (2017).

indicates that 9 of the 10 *p*-values are < .01. This percentage is consistent with the 10 primary studies having considerable evidential value. More exactly, the extreme heaping of *p*-values near 0 reflects some combination of adequate sample sizes in the primary studies and an appreciable true difference between the effect of L1 glossing and the effect of L2 glossing for learners ranging approximately from elementary to intermediate in proficiency.

Still focusing on Figure 6 we now come to the dashed curve, which becomes relevant if the observed *p*-curve has minimal right skew. This dashed curve represents the approximate minimum of right skew consistent with substantively, or pedagogically, significant evidential value. The corresponding significance tests (i.e., tests 4 to 6) are, respectively, a binomial test and two Stouffer's tests. If they show $p > .05$, it is warranted to conclude that the observed *p*-curve represents a modicum of evidential value. For the *p*-curve shown in Figure 6, these three tests give $p > .999$. If they had yielded $p < .05$, that would have indicated either the likely absence of a pedagogically significant true effect or an insufficient level of statistical power to detect whatever true effect may be present. Lack of power could be corrected by finding additional relevant primary studies. It would be an especially bad sign if the left end of the solid line remains below the dashed line. Conversely, *p*-hacking would be fairly dramatically indicated if the right end of the solid line *rises* above the dashed line. Technical details about the *p*-curve app and its output are available both at http://www.p-curve.com/ and in material that accompanies a computed *p*-curve graph.

Finally, the small print at the top of Figure 6 gives 96% as the estimate of the average "observed," or post hoc, statistical power of the 10 studies underlying the *p*-curve. In most

research situations observed power, unlike *a priori* power, is of little or no interest for reasons set out by Lakens (2021). The textual output of the *p*-curve app explains the special role of observed power with respect to the app's graphical output.

### *A FURTHER STEP IN P-CURVE ANALYSIS: CHECKING ROBUSTNESS*

At this point in a *p*-curve analysis, an important question remains: Is the observed *p*-curve overly dependent on outliers—that is, on primary studies that yielded extremely low or extremely high significant *p*-values? This question can be answered by deleting *p*-values at either extreme of an observed *p*-curve to see how the curve changes. Technically, a type of "robustness test" is called for. Here again the *p*-curve app employs Stouffer's test. In the app's robustness testing, one implementation of Stouffer's test focuses on the observed 0↔.05 *p*-curve. In the present case that test continued to yield $p < .05$ at every stage of the value-by-value cumulative omission of the four lowest *p*-values. A second robustness test continued to yield $p < .05$ when the procedure was repeated, but this time working from the highest *p*-value downward. A third robustness test focusing on the 0↔.025 half-curve consistently yielded $p < .05$ when working from the lowest *p*-value up. It can therefore be concluded that the 10 studies involved in the *p*-curve analysis have robust evidential value.

### *DISCUSSION*

From the results given in Table 4 and from the shape of the *p*-curve shown in Figure 6 it is inferable, first, that the 10 primary studies covered by the main *p*-curve analysis have notable evidential value. It is inferable also that there was no worrying *p*-hacking, there being no *p*-values between .02 and .05. Moreover, the results of the robustness tests are reassuring. All this matters and is encouraging because the effect sizes corresponding to the 10 underlying *p*-values will have made a prominent contribution to the estimate of average effect size that Kim et al. (2020) reported for the research hypothesis at issue in this case study. However, the proportional contribution of present results must be smaller with respect to the estimate of Yanagisawa et al. (2020) because that meta-analysis focused on studies of incidental learning and because it included many unpublished studies not covered by this case study.

Although the previously mentioned results of the *p*-curve analysis are encouraging, it should be borne in mind that *p*-curve analysis is not a tool for assessing overall study quality: Studies should be vetted for quality *before* *p*-curving begins. To say more about quality, recall that 4 of the 10 primary studies figuring in this case study were quasi-experimental. It is not controversial to say that results from quasi-experiments tend to be less credible than results from experiments. It is noteworthy also that each of the 10 primary studies generated raw test scores that were binary (i.e., correct or incorrect) for each target vocabulary item and that all of the researchers used ANOVA and/or *t*-tests for which the binary scores were subtotaled by learner. It is now widely recognized that it is usually preferable to analyze binary vocabulary test scores by using mixed-effects logistic regression (Linck & Cunnings, 2015). See Boers (in press) for insightful discussion of the quality of published studies of the effects of glossing on L2 vocabulary learning.

Lastly, in an earlier section it was mentioned that Simonsohn et al. (2014a) suggested 10 as the minimum number of primary studies needed for a useful *p*-curve if statistical power is adequate (Simonsohn et al., 2014a). It was mentioned too that Simonsohn et al. based their estimate on a mean sample size of 20 participants. The primary studies covered by the present case study had a median sample size of 52 participants (i.e., learners). Consequently, it is likely that 10 studies were indeed sufficient for this study. Still, it would have been preferable to have had *p*-values from additional studies. The aforementioned study by Kazerouni and Rassaei (2016) was not included in the present *p*-curve analysis because it did not meet the main criterion of the preset inclusion rule, namely, coverage by Kim et al. (2020) or Yanagisawa et al. (2020). However, application of Welch's *t*-test to Kazerouni and Rassaei's immediate posttest scores produced $t = 5.8437$, $df = 77.156$, $p = .0000001$. This result is consistent with L1 glossing being comparatively effective for lowish proficiency learners like those in Kazerouni and Rassaei's study. Interested readers may extend the 10-study *p*-curve analysis by adding *t(77.156)=5.8437* to the list of statistics given in Table 3 and then submitting the extended list to the *p*-curve app (http://www.p-curve.com).

## TWO ALTERNATIVE STRATEGIES IN *P*-CURVING

Until relatively recently quantitative researchers of SLA have relied extremely heavily on a small number of inferential analytic procedures, especially ANOVA, two-sample *t*-tests, and two-sample rank-based tests such as the WMW test (Lindstromberg, 2016; Plonsky, 2013; Plonsky & Gass, 2011). With respect to these procedures, the statistical significance tests implemented by the *p*-curve app are able to warrant the conclusion that a *p*-curve reflects appreciable evidential value even when a minority of the *p*-values happen to reflect *p*-hacking—including some bold *p*-hacking. However, it was asserted by Head et al. (2015) that the app's significance tests may lack sensitivity to the presence of modest *p*-hacking, even though modest *p*-hacking is presumably more common than bold *p*-hacking. Head et al. stated in particular that the approach taken by the *p*-curve app can result in signs of modest *p*-hacking being masked when a strong average true effect results in dramatic heaping of *p*-values near zero. As a means of detecting modest *p*-hacking in this circumstance Head et al. recommended use of a one-sided sign test (a version of the binomial test). For this test the *H0* would be that a specified segment of a *p*-curve is either right skewed or uniform. So, $p < .05$ would indicate *left* skew. The sign test could, for example, focus on the range .02↔.05 by comparing the number of *p*-values in the subrange .02↔.035 with the number in the subrange .035↔.05. Obviously, running such a test would only make sense if there were enough *p*-values in the full range of interest.

If the recommendation of Head et al. (2015) is not applicable due to a dearth of relevant *p*-values, a practical alternative is suggested by best practice in attempting to understand a correlation. That is, interpretation should not cease with consideration of the correlation coefficient and perhaps of the associated *p*-value as well. Rather, it is always advisable to examine the corresponding scatterplot. This is the lesson afforded by "Anscombe's quartet" (Anscombe, 1973). Similarly, *p*-curvers should not place so much weight on the results of significance tests (e.g., the ones carried out by the *p*-curve app) that they

overlook clues visibly afforded by the relevant *p*-curve graph. The next section deals with a case in point.

## AN AUTHENTIC *P*-CURVE SHOWING RIGHT SKEW AND LEFT SKEW

It may now be instructive to compare the *p*-curve shown in Figure 6 with the one shown in Figure 7. The latter is a simplified but essentially accurate rendering of a *p*-curve graph from the *p*-curve app that was displayed by Vitta and Al-Hoorie (2020) in relation to their meta-analytic review of studies having to do with "the flipped classroom" in SLA. This new *p*-curve graph is presented here to illustrate certain issues that may arise when a *p*-curve is to be interpreted.

Two key facts regarding Figure 7 are as follows. First, this figure is based on all 45 of the significant *p*-values from Vitta and Al-Hoorie's (2020) total sample of primary studies ($k = 56$). Second, those studies addressed at least three research hypotheses. A consequence of this last fact is that Figure 7 resists interpretation because it does not show which *p*-values in the curve relate to which research hypothesis. To resolve the unclarity it would be necessary to construct a separate *p*-curve for each research hypothesis (Simonsohn et al., 2014a). A second problem is that Vitta and Al-Hoorie made it clear that observed effects of flipped instruction have been heterogeneous with respect to each of the research hypotheses that they took into account. But, as mentioned in an earlier section of this article, it can be difficult to interpret a 0↔.05 *p*-curve when the underlying effects are heterogeneous.

A third issue regarding Figure 7 has to do with the information it would have conveyed (a) if all the underlying studies had addressed a single research hypothesis and (b) if it were reasonable to assume that the relevant observed effects correspond to an approximately

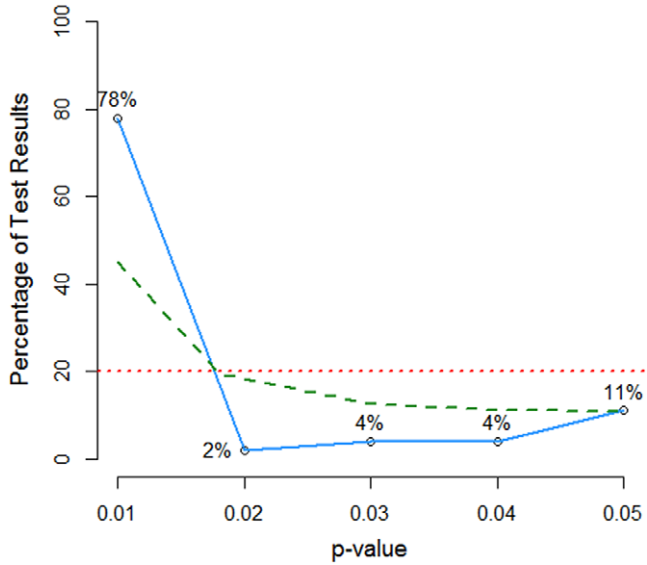


FIGURE 7. The key elements of the *p*-curve presented by Vitta and Al-Hoorie (2020).

homogeneous population effect. Given these two conditions, the significance tests which the *p*-curve app carried out on Vitta and Al-Hoorie's (2020) data will indeed have indicated that the underlying *p*-values have considerable evidential value. This is inferable from the version of Figure 7 given in Vitta and Al-Hoorie's article and from the associated commentary. Note in particular that Figure 7 shows that 78% of the underlying *p*-values are below .01. Nevertheless, their *p*-curve is skewed *left* between .02 and .05. For instance, only 2% of the *p*-values lie between .01 and .02 whereas 11% lie between .04 and .05. This is a classic warning sign: In the absence of *p*-hacking, we would expect the .01↔.05 section of the *p*-curve to be either flat or slightly *right* skewed. It is therefore surprising that Vitta and Al-Hoorie stated that this *p*-curve furnishes no evidence that there may have been *p*-hacking. Had a separate *p*-curve been calculated for each main research hypothesis, it could then be seen, for instance, whether any indication of *p*-hacking is confined to one hypothesis in particular.

When significance tests indicate that a set of *p*-values have evidential value but the corresponding *p*-curve graph indicates concurrent modest *p*-hacking, neither indication should be ignored. True, the results of the significance tests may warrant bolstered confidence that the relevant stream of research is worth pursuing. However, observed left skew may suggest that certain studies should be re-read or even that the authors should be asked for additional information or for the raw data. Ultimately it may be prudent to class one or more studies as dubious.

## OVERALL SUMMARY

The foregoing sections of this article have outlined the factual basis, the rationale, and the key steps of *p*-curving for the purpose of gaining an informed impression of the degree to which a set of *p*-values reflects *p*-hacking. It was stressed that a *p*-curve analysis should be based on significant *p*-values relating to the same research hypothesis. Also stressed was the importance of transparency in reporting decisions with potential to affect the outcome of a *p*-curve analysis. This transparency should include, for example, an account of decisions made about which *p*-values should or should not be included in the analysis at hand.

## CONCLUSION

Currently, there appear to be no serious competitors to *p*-curving as a means of gaining an actionable impression of the probability that a stream of research has been affected by *p*-hacking. Fortunately, *p*-curve analyses are comparatively easy to carry out. This is especially true with respect to a research stream that a prospective *p*-curver knows well. *P*-curving software is freely available; information about its use is easy to find at http://www.p-curve.com/ and elsewhere (e.g., Head et al., 2015). Moreover, good examples of how to report a *p*-curve analysis are becoming more numerous (e.g., Dick et al.; 2019; Lakens, 2018; Vogel & Homberg, 2020).

Lastly, if it became routine for SLA researchers to carry out and report *p*-curve analyses, it would likely become evident far sooner than would otherwise be the case whether or to what extent *p*-hacking is a problem in some areas of SLA quantitative research.

## SUPPLEMENTARY MATERIALS

To view supplementary material for this article, please visit http://dx.doi.org/10.1017/S0272263121000516.

## NOTES

[1]Gelman and Loken (2013) acknowledged that exploration does have its place in data analysis. Their position is that a study that included exploratory data analysis should be followed up with a direct replication, and that both studies should be reported in the same paper.

[2]In discussions of methods for detecting file-drawering it is routine for authors to misleadingly speak of these methods as methods for detecting "publication bias" (e.g., Carter et al., 2019; van Aert et al., 2019).

[3]Van Aert and van Assen (2021) recommended against using trim and fill even for detecting publication bias (by which they meant file-drawering) due to the tendency of trim and fill to give misleading results in common situations, such as when true effect sizes vary across contexts. Carter et al. (2019, p. 121) concluded that trim and fill "tends to be outperformed by other methods and generally fails as heterogeneity increases."

[4]There are exceptions (e.g., Gelman, 2018). However, some of these exceptions are unlikely to be problematic in SLA quantitative studies, particularly in studies involving ANOVA and *t*-tests. For example, when the effect size is 0, a *p*-curve based on results from Welch's IS *t*-test may have a slightly more ragged outline than would be expected if Student's IS-*t*-test were used (Figure 1). Bishop and Thompson (2016) described an additional exception involving studies with multiple outcome variables.

[5]Currently there are two *p*-curve methods for effect size estimation that do take all *p*-values into account, the method of Hedges (1992) and the method of van Aert and van Assen (2021). When there has been moderate *p*-hacking, these methods perform better than standard random effects meta-analysis (Carter et al., 2019; van Aert & van Assen, 2021). Hedge's *p*-curve method for estimating effect sizes and the method of van Aert and van Assen can be implemented by using, respectively, the R packages *weightr* (Coburn & Vevea, 2016) and *puniform* (van Aert, 2021).

[6]Also on page 453 (note 17) Simonsohn et al. (2014a) state the rationale for the correct choice of *p*-values for a reversing interaction.

[7]Yanagisawa et al. (2020) found no persuasive evidence of an interaction between language and proficiency. Kim et al. (2020) reached the opposite conclusion.

## PRIMARY STUDIES INCLUDED IN CASE STUDY

Arpaci, D. (2016). The effects of accessing L1 versus L2 definitional glosses on L2 learners' reading comprehension and vocabulary learning. *Eurasian Journal of Applied Linguistics*, *2*, 15–29. https://www.ejal.info/index.php/ejal/issue/view/10

Ertürk, Z. (2016). The effect of glossing on EFL learners' incidental vocabulary learning in reading. *Procedia: Social and Behavioral Sciences*, *232*, 373–381. https://doi.org/10.1016/j.sbspro.2016.10.052

Farvardin, M., & Biria, R. (2012). The impact of gloss types on Iranian EFL students' reading comprehension and lexical retention. *International Journal of Instruction*, *5*, 99–114. http://www.e-iji.net/volumes/317-january-2012-volume-5-number-1

Ko, M.-H. (1995). Glossing in incidental and intentional learning of foreign language vocabulary and reading. *University of Hawai'i Working Papers in ESL*, *13*, 49–94. https://scholarspace.manoa.hawaii.edu/handle/10125/40761

Ko, M.-H. (2017). The relationship between gloss type and L2 proficiency in incidental vocabulary learning. *Modern English Education*, *18*, 47–69. http://www.dbpia.co.kr/Article/NODE07255877

Mitarai, Y., & Aizawa, K. (1999). The effects of different types of glosses in vocabulary learning and reading comprehension. *ARELE: Annual Review of English Language Education in Japan*, *10*, 73–82.

Öztürk, M., & Yorgancı, M. (2017). Effects of L1 and L2 glosses on incidental vocabulary learning of EFL prep students. *Turkish Studies: International Periodical for the Languages, Literature and History of Turkish or Turkic*, *12*, 635–656. http://dx.doi.org/10.7827/TurkishStudies.11432

Pishghadam, R., & Ghahari, S. (2011). The impact of glossing on incidental vocabulary learning: A comparative study. *Iranian EFL Journal*, *7*, 8–29. https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.965.6528&rep=rep1&type=pdf

Shiki, O. (2008). Effects of glosses on incidental vocabulary learning: Which gloss-type works better, L1, L2, single choice, or multiple choices for Japanese university students? *Journal of Inquiry and Research*, *87*, 39–56. http://doi.org/10.18956/00006209

Yoshii, M. (2006). L1 and L2 glosses: Their effects of incidental vocabulary learning. *Language Learning & Technology*, *10*, 85–101. https://www.lltjournal.org/item/2563

## REFERENCES

Anscombe, F. (1973). Graphs in statistical analysis. *The American Statistician*, *27*, 17–21. https://doi.org/10.2307/2682899

Arnholt, A. (2017). R package *BSDA* (Basic statistics and data analysis), Version 1.20. (Computer freeware). https://www.rdocumentation.org/packages/BSDA/versions/1.2.0

Bakker, M., & Wicherts, J. (2011). The (mis) reporting of statistical results in psychology journals. *Behavior Research Methods*, *43*, 666–678. https://doi.org/10.3758/s13428-011-0089-5

Bakker, M., & Wicherts, J. (2014). Outlier removal, sum scores, and the inflation of the type I error rate in independent samples t tests: The power of alternatives and recommendations. *Psychological Methods*, *19*, 409–427. https://doi.org/10.1037/met0000014

Barcroft, J. (2015). *Lexical input processing and vocabulary learning*. John Benjamins.

Bishop, D., & Thompson, P. (2016). Problems in using *p*-curve analysis and text-mining to detect rate of *p*-hacking and evidential value. *PeerJ*, *4*:e1715. https://doi.org/10.7717/peerj.1715

BITSS (Berkeley Initiative for Transparency in the Social Sciences). (2017). P-curve: A tool for detecting publication bias. https://www.bitss.org/p-curve-a-tool-for-detecting-publication-bias/

Boers, F. (in press). Glossing and vocabulary learning. *Language Teaching*. https://doi.org/10.1017/S0261444821000252

Borenstein, M., Hedges, L., Higgins, J., & Rothstein, H. (2009). *Introduction to meta-analysis*. Wiley.

Carter, E., Schönbrodt, F., Gervais, W., & Hilgard, J. (2019). Correcting for bias in psychology: A comparison of meta-analytic methods. *Advances in Methods and Practices in Psychological Science*, *2*, 115–144. https://doi.org/10.1177/2515245919847196

Chambers, C. (2017). *The seven deadly sins of psychology: A manifesto for reforming the culture of scientific practice*. Princeton University Press.

Coburn, K., & Vevea, J. (2016). weightr: Estimating weight-function models for publication, Version 2.0.2. (Computer freeware). https://CRAN.R-project.org/package=weightr

Dick, A., Garcia, N., Pruden, S., Thompson, W., Hawes, S., Sutherland, M., Riedel, M., Laird, A., & Gonzalez, R. (2019). No evidence for a bilingual executive function advantage in the ABCD study. *Nature Human Behavior*, *3*, 692–701. https://doi.org/10.1038/s41562-019-0609-3

Duval, S., & Tweedie, R. (2000). Trim and fill: A simple funnel-plot–based method of testing and adjusting for publication bias in metaanalysis. *Biometrics*, *56*, 455–463. https://doi.org/10.1111/j.0006-341X.2000.00455.x

Egger, M., Smith, G., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *British Medical Journal*, *315*, 629–634. https://doi.org/10.1136/bmj.315.7109.629

Erdfelder, E., & Heck, D. (2019). *P*-curve: A word of caution. *Zeitschrift für Psychologie*, *227*, 249–260. https://doi.org/10.1027/a000001

Fidler, F., & Wilcox, J. (2018). Reproducibility of scientific results. *Stanford encyclopedia of philosophy*. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/entries/scientific-reproducibility/

Gelman, A. (2018). The *p*-curve, *p*-uniform, and Hedges (1984). Methods for meta-analysis under *p-hacking*: An exchange with Blake McShane, Uri Simosohn, and Marcel van Assen. *Stat modeling, causal inference, and social science*, 26 February. https://statmodeling.stat.columbia.edu/2018/02/26/p-curve-p-uniform-hedges-1984-methods-meta-analysis-selection-bias-exchange-blake-mcshane-uri-simosohn/

Gelman, A., & Loken, E. (2013). *The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time*. Department of Statistics, Columbia University. http://www.stat.columbia.edu/~gelman/research/unpublished/p_hacking.pdf

Hartgerink, C. (2017). Reanalyzing Head et al. (2015): Investigating the robustness of widespread *p*-hacking. *PeerJ Preprints*, *5*, e3068. https://doi.org/10.7717/peerj.3068

Head, M., Holman, L., Lanfear, R., Kahn, A., & Jennions, M. (2015). The extent and consequences of *p*-hacking in science. *PLOS Biology*, *13*, e1002106. https://doi.org/10.1371/journal.pbio.1002106

Hedges, L. (1992). Modeling publication selection effects in meta-analysis. *Statistical Science*, *7*, 246–255. https://projecteuclid.org/euclid.ss/1177011364

John, L., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science*, *23*, 524–532. https://doi.org/10.1177/0956797611430953

Kazerouni, Z., & Rassaei, E. (2016). The effects of L1 and L2 glossing on the retention of L2 vocabulary in intentional and incidental settings. *Journal of Studies in Learning and Teaching English*, *5*, 119–150. http://jslte.iaushiraz.ac.ir/issue_112611_112618.html

Kim, H., Lee, J., & Lee, H. (2020). The relative effects of L1 and L2 glosses on L2 learning: A meta-analysis. *Language Teaching Research*. Advance view. https://doi.org/10.1177/1362168820981394

Lakens, D. (2014). What *p*-hacking really looks like: A comment on Masicampo & Lalande. (2012). *Quarterly Journal of Experimental Psychology A*, *68*, 829–832. https://doi.org/10.1080/17470218.2014.982664

Lakens, D. (2018). Professors are not elderly: Evaluating the evidential value of two social priming effects through *p*-curve analyses. Eindhoven University of Technology. https://psyarxiv.com/3m5y9/

Lakens, D. (2021). Sample size justification. *PsyArXiv* https://psyarxiv.com/9d3yf/

Lakens, D., Scheel, A., & Isager, P. (2018). Equivalence testing for psychological research. *Advances in Methods and Practices in Psychological Science*, *1*, 1259–69. https://doi.org/10.1177/2515245918770963

Light, R., & Pillemer, D. (1984). *Summing up: The science of reviewing research*. Harvard University Press.

Linck, J., & Cunnings, J. (2015). The utility and application of mixed-effects models in second language research. *Language Learning*, *65*, 185–207. https://doi.org/10.1111/lang.12117

Lindstromberg, S. (2016). Inferential statistics in *Language Teaching Research*: A review and ways forward. *Language Teaching Research*, *20*, 741–768. https://doi.org/10.1177/1362168816649979

McShane, B., Böckenholt, U., & Hansen, K. (2016). Adjusting for publication bias in meta-analysis: An evaluation of selection methods and some cautionary notes. *Perspectives on Psychological Science*, *11*, 730–749. https://doi.org/10.1177/1745691616662243

Norris, J. (2015). Statistical significance testing in second language research: Basic problems and suggestions for reform. *Language Learning*, *65*, 97–126. https://doi.org/10.1111/lang.12114

Plonsky, L. (2013). Study quality in SLA: An assessment of designs, analyses, and reporting practices in quantitative L2 research. *Studies in Second Language Acquisition*, *35*, 655–687. https://doi.org/10.1017/S0272263113000399

Plonsky, L., & Gass, S. (2011). Quantitative research methods, study quality, and outcomes: The case of interaction research. *Language Learning*, *61*, 325–366. https://doi.org/10.1111/j.1467-9922.2011.00640.x

Plonsky, L. & Oswald, F. (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., Sudina, E., & Hu, Y. (2021). Applying meta-analysis to research on bilingualism: An introduction. *Bilingualism: Language and Cognition*. Advance online publication. https://doi.org/10.1017/S1366728920000760

Pollet T., & van der Meij, L. (2017). To remove or not to remove: The impact of outlier handling on significance testing in testosterone data. *Adaptive Human Behavior and Physiology*, *3*, 43–60. https://doi.org/10.1007/s40750-016-0050-z

Roettger, T. (2019). Researcher degrees of freedom in phonetic research. *Laboratory Phonology: Journal of the Association for Laboratory Phonology*, *10*, 1–27. https://doi.org/10.5334/labphon.147

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, *86*, 638–641. https://doi.org/10.1037/0033-2909.86.3.638

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

Simmons, J., Nelson, L., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*, 1359–1366. http://doi.org/10.1177/0956797611417632

Simonsohn, U., Nelson, L., & Simmons, J. (2014a). *P*-curve: A key to the file drawer. *Journal of Experimental Psychology: General*, *143*, 534–547. http://dx.doi.org/10.1037/a0033242

Simonsohn, U., & Nelson, L., & Simmons, J. (2014b). *P*-curve and effect size: Correcting for publication bias using only significant results. *Perspectives on Psychological Science*, *9*, 666–681. https://doi.org/10.1177/1745691614553988

Simonsohn, U., Simmons, J., & Nelson, L. (2015). Better *p*-curves: Making *p*-curve analysis more robust to errors, fraud, and ambitious *p*-hacking, A reply to Ulrich and Miller (2015). *Journal of Experimental Psychology: General*, *144*, 1146–1152. http://dx.doi.org/10.1037/xge0000104

Simonsohn, U., Nelson, L., & Simmons, J. (2017). *P-curve app 4.06*. (Computer freeware). http://www.p-curve.com/app4/

Simonsohn, U., Nelson, L., & Simmons, J. (2019). *P*-curve won't do your laundry, but it will distinguish replicable from non-replicable findings in observational research: Comment on Bruns & Ioannidis (2016). *PLoS ONE 14*, e0213454. https://doi.org/10.1371/journal.pone.0213454

van Aert, R. (2021). puniform: Meta-analysis methods correcting for publication bias, Version 0.2.4. (Computer freeware). https://github.com/RobbievanAert/puniform

van Aert, R., & van Assen, M. (2021). Correcting for publication bias in a meta-analysis with the *p-uniform** method. Open Science Framework. https://doi.org/10.31222/osf.io/zqjr9

van Assen, M., van Aert, R., & Wicherts, J. (2015). Meta-analysis using effect size distributions of only statistically significant studies. *Psychological Methods*, *20*, 293–309. http://dx.doi.org/10.1037/met0000025

van Aert, R., Wicherts, J., & van Assen, M. (2019). Publication bias examined in meta-analyses from psychology and medicine: A meta-meta-analysis. *PLoS ONE*, *14*, e0215052. https://doi.org/10.1371/journal.pone.0215052

Vitta, J., & Al-Hoorie, A. (2020). *The flipped classroom in second language learning: A meta-analysis*. Language Teaching Research. Advance view. https://doi.org/10.1177/1362168820981403

Vogel, D., & Homberg, F. (2020). *P*-hacking, *p*-curves, and the PSM–performance relationship: Is there evidential value? *Public Administration Review*, *81*, 191–204. http://dx.doi.org/10.1111/puar.13273

Westfall, J. (2016). Five different "Cohen's *d*" statistics for within-subject designs. *Cookie Scientist: Designing experiments and analyzing data*. 25 March. http://jakewestfall.org/blog/index.php/2016/03/25/five-different-cohens-d-statistics-for-within-subject-designs/

Wicherts, J., Veldkamp, C., Augusteijn, H., Bakker, M., van Aert, Robbie, M., & van Assen, M. (2016). Degrees of freedom in planning, running, analyzing, and reporting psychological studies: A checklist to avoid *p*-hacking, *Frontiers in Psychology*, *7*, 1832. https://www.frontiersin.org/article/10.3389/fpsyg.2016.01832

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