UNIVERSITY of CALIFORNIA PRESS | Collabra: Psychology

Boustani, S., & Shanks, D. R. (2022). Heterogeneity and Publication Bias in Research on Test-Potentiated New Learning. *Collabra: Psychology*, *8*(1). https://doi.org/10.1525/collabra.31996

Cognitive Psychology

Heterogeneity and Publication Bias in Research on Test-Potentiated New Learning

Shaun Boustani[1] [a], David R. Shanks[1]

[1] Experimental Psychology, University College London, UK

Keywords: publication bias, meta-analysis, testing effect, test-potentiated new learning

https://doi.org/10.1525/collabra.31996

Collabra: Psychology

Vol. 8, Issue 1, 2022

Prior retrieval practice potentiates new learning. A recent meta-analysis of this test-potentiated new learning (TPNL) effect by Chan, Meissner, and Davis (2018) concluded that it is a robust and reliable finding (Hedges' g = 0.44). Although Chan et al. discussed three different experimental designs that have been employed to study TPNL, we argue that their meta-analysis failed to adequately distinguish the findings from these different designs, acknowledge the significance of the substantial between-study heterogeneity across all pooled effects, and assess the degree of publication bias in the sample. We conducted a new meta-analysis that assessed the designs separately and applied appropriate corrections for publication bias. We found that studies using a standard design yield weak evidence of a TPNL effect, studies using pre-testing yield a small but reliable effect, and studies using interleaving designs yield weak evidence of a negative effect. Compared to Chan et al.'s conclusions, these reanalyses cast TPNL in a very different light and point to a pressing need for preregistered experiments to assess its reproducibility in the absence of publication bias.

Test-potentiated new learning (TPNL) is the finding that prior retrieval practice potentiates new learning. For example, Szpunar, Khan, and Schacter (2013) had participants study an introductory statistics video divided into four segments. After each of segments 1-3, participants in a retrieval group took a test (without corrective feedback) about the preceding segment while those in a control group were not tested. After studying segment 4, both groups were tested. Szpunar et al. found that participants who had undertaken prior retrieval practice recalled far more of the content of segment 4 than those who had not. Thus, retrieval practice can enhance subsequent learning of new material. Like the classic testing effect, in which tests consolidate the information retrieved in the test (McDermott, 2021; Roediger & Karpicke, 2006; Yang et al., 2021), TPNL has important implications and suggests that testing can play a useful role in educational settings.

A recent meta-analysis by Chan, Meissner, and Davis (2018) concludes, and seemingly confirms, that TPNL is a robust and reliable finding (Hedges' g = 0.44). However, Chan et al. did not adequately assess the extent of publication bias nor address the substantial between-study heterogeneity in their sample. As we demonstrate, these limitations make this conclusion unwarranted and raise a question mark about the robustness of TPNL. We show that the aggregated effect needs to be divided into meaningful sub-components, and that doing so changes the conclusions that can be drawn.

In their meta-analysis Chan et al. pooled data from a heterogenous and diverse sample of studies using different materials, testing formats, and populations. This is an appropriate practice for an exploratory analysis focused on the broad impact of retrieval practice, and a random-effects model was used, which does not assume homogeneity in effect sizes (Borenstein, 2009). Nevertheless, the pooling resulted in substantial between-study heterogeneity (I^2 = 86%) which was not explained by the subgroup analyses they conducted. In this comment we show that Chan et al.'s pooling was unjustified and that their total (pooled) sample comprised three distinct subsets. By disaggregating these effects, we substantially reduce heterogeneity and show that they are probably underpinned by different cognitive mechanisms. Furthermore, Chan et al. also failed to test, or correct for, publication bias in their meta-analysis. We present evidence that these three subsets have very different aggregate effects sizes when corrected for publication bias. Indeed, we show that the magnitude of two of these effects may not be different from zero.

Publication bias in Chan et al.'s (2018) meta-analysis

In the Chan et al. (2018) meta-analysis, the only assessment of publication bias came from comparing the effect

a shaun.boustani@sydney.edu.au

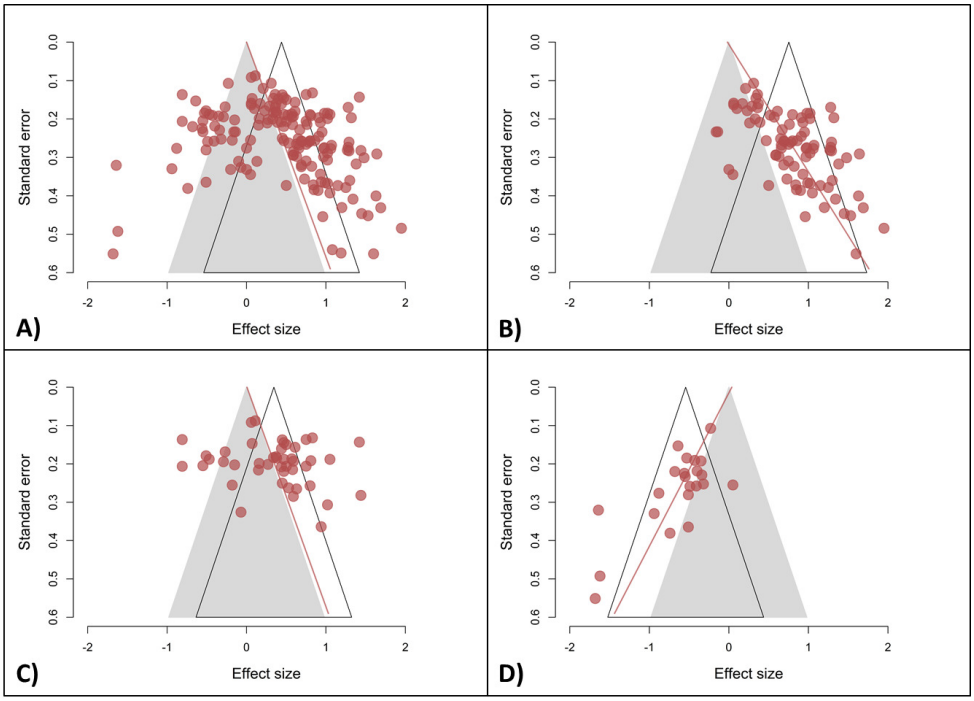

Figure 1. Funnel plots of effect sizes (Hedges' *g*) in reanalyses of Chan et al.'s (2018) dataset

The red points are study estimates and the red line depicts Egger's regression test. The gray triangle is centered on an effect size of zero and represents the space in which individual study effects would not be statistically significant at $p < .05$. The open triangle is centered on the meta-analytic effect size estimate. A: Overall sample. B: Standard sample. C: Pre-testing sample. D: Interleaved test sample.

sizes of published and unpublished data. This assessment yielded a significant difference ($g = 0.50$ for published studies versus $g = 0.19$ for unpublished ones). However, instead of interpreting this as suggestive of publication bias (which would need correcting), Chan et al. proposed that it arose because publication status and experimental design tended to be confounded. But rather than respond to this confounding by conducting alternative or more targeted publication bias tests, Chan et al. simply assumed that publication bias was not a concern in their meta-analysis. We argue that this was inadequate and additional tests are needed.

There are many other ways of testing for publication bias. A classic method is to examine a funnel plot (Figure 1, panel A) of the data and apply Egger's test (Egger et al., 1997) to determine whether the experiments' effect sizes are correlated with their standard errors. Evidence that studies with larger standard errors yield larger effect sizes is taken as evidence of publication bias. As standard error is dependent on sample size, it is taken as a metric of study precision, which is the reliability of the effect size estimate. A significant positive correlation implies that more precise studies, those with smaller standard errors and larger sample sizes, are associated with smaller effects. This means that the aggregated effect size likely has small study bias. Many methods have been proposed to adjust for bias, as discussed further below.

To apply Egger's test, we first replicated Chan et al.'s (2018) meta-analysis, reanalyzing their sample available though OSF (https://osf.io/7wufc/). We applied a random-effects (RE) model (see Hedges & Vevea, 1998), using R with the *meta*, *metafor*, and *dmetar* packages. We made two minor modifications by specifying the Paule-Mandel estimator of τ^2 (Paule & Mandel, 1982) and the Hartung-Knapp-Sidik-Jonkman (HKSJ) (Hartung & Knapp, 2001; Sidik & Jonkman, 2002) adjustment method. This analysis generated an aggregate effect point estimate identical to that observed by Chan et al., with near identical 95% CIs and *Q* value, $g = 0.44$, CI [0.34, 0.54], $t(158) = 8.46$, $p < .001$. This confirms that their results are reproducible. As in Chan et al., the analysis also detected significant heterogeneity, $Q(158) = 1133.94$, $p < 0.001$, with a substantial amount due to between-study variation in effect sizes, $I^2 = 86\%$. That is, the variation was not only due to sampling error, but likely due to significant variations in true effects. Further analyses found that this substantial heterogeneity was also confirmed by *tau* ($\tau = 0.60$). Tau is the estimated standard deviation of the true effects in the sample and quantifies dispersion (Borenstein, 2009). In our sample, the tau score indicates that the effects span a very wide range of effect sizes from -0.76 to 1.64, the range incorporating 95% of effects. So, although the aggregated effect was positive and significant, there is very wide variation that needs to be explained as well as a substantial number of studies finding that testing impairs new learning. As discussed below, we believe this is likely due to improper aggregation of methodologies.

Applying Egger's method to this dataset reveals substantial asymmetry, $B = 1.85$, $t(158) = 3.16$, $p = .002$, which is usually interpreted as a potential signal of publication bias. To select an appropriate test, we followed the principled approach developed by Carter et al. (2019) and its application (http://www.shinyapps.org/apps/metaExplorer/). We specified severity of publication bias and the extent of questionable research practices as 'medium', heterogeneity (τ) = 0.2

(note that this parameter reflects the assumed true level of heterogeneity, not the amount observed in the RE analysis), the true effect size as 0.5, and the number of studies in the meta-analysis = 100 (the largest number possible). Good performance was defined as a maximum false positive rate = 20%. Under these conditions, 3PSM and PET provide suitable estimates. Other well-known correction methods such as Trim-and-Fill and *p*-curve generate unacceptable maximum false positive rates. For example, under these plausible conditions, Trim-and-Fill has a false positive rate of 100% when there is no true effect. The 3PSM and PET methods were also acceptable when the number of studies was set at 20, 40, and 60. This makes them suitable for assessing bias in the interleaving, pre-testing, and standard designs, respectively (see below).

When the bias was adjusted for using the three-parameter selection model (3PSM; Iyengar & Greenhouse, 1988; Vevea & Hedges, 1995; Vevea & Woods, 2005), the overall TPNL effect estimate was entirely eliminated and the aggregate effect size was no longer reliable, g = 0.02, CI [-0.14, 0.18]), $Z(158) = 0.22$, $p = .82$. A likelihood ratio test confirmed that the 3PSM adjusted model fit the data significantly better than an unadjusted model, $\chi^2(1) = 42.90$, $p < 0.0001$. This null effect was also found when bias was adjusted for using the Precision-Effect Test (PET; Stanley, 2008), g = -0.02, CI [-0.26, 0.22]), $t(158) = 0.88$, $p = .88$. This test takes the intercept of the Egger regression when the standard error is zero as the best estimate of the true effect (illustrated by the red line in Figure 1A). These are striking results which question the existence of a robust benefit of testing on new learning.

One important factor in this assessment of publication bias, however, is the large amount of between-study heterogeneity observed in Chan et al.'s (2018) meta-analysis, which could impact the accuracy of publication bias assessment (Peters et al., 2010; Terrin et al., 2003). As we show next, when the TPNL effect is separated based on testing procedure, the bias-adjusted effect size estimates are very different.

Heterogeneity in Chan et al.'s (2018) meta-analysis

As discussed, the Chan et al. (2018) meta-analysis found substantial between-study heterogeneity. Although this is expected when data are pooled from studies using different experimental designs, materials, and participant samples, such large heterogeneity questions the validity of pooling these effects. Rather than quantifying a single true effect, the Chan et al. (2018) meta-analysis may have combined several distinct true effects.

The most obvious pooling that could have resulted in the observed heterogeneity was across the experimental designs noted by Chan et al. (2018): the *standard*, *pre-testing*, and *interleaved* designs. Full details are given in Chan et al., but in brief the differences are as follows. In standard designs, which include the Szpunar et al. (2013) study described above, participants study an initial set of materials and then complete an interval task, usually either a test or restudy. After this, they study a new set of materials and then complete a test. New learning is classified as learning of the new set of materials, and the second test is the criterial assessment. In a pre-testing design, participants do not study a set of materials first, but rather complete an initial test and are then presented with the materials, after which they complete a second test for memory of those materials. In that respect, pre-testing does not require episodic retrieval and does not feature multiple lists of materials. In an interleaved design, rather than new materials being presented separately to original materials, the new and original materials are interleaved. For example, in Davis and Chan's (2015) experiments, participants first studied a face-name pair, and then either restudied the pair, or were tested. In the testing condition, participants were presented with the face and retrieved the name. Crucially, directly after the interval task, participants were provided an additional piece of information, a profession which was presented alongside the original face-name pair (i.e., participants studied a face-name-profession triad). After this, participants completed a criterial test where they were presented the face, but now had to retrieve the profession. As we will discuss below, it is likely that the effects produced by these designs are underpinned by different cognitive mechanisms.

One important distinction between the current analyses and those performed by Chan et al. (2018) is our treatment of the interleaved study design. In the Chan et al. meta-analytic sample, some pre-testing and interleaved studies were confounded. In their coding scheme pre-testing studies in which new learning was provided as feedback were coded as also having used an interleaved design (for example, Knight et al., 2012; Kornell et al., 2009; Vaughn et al., 2016). Although this is defensible as the initial test and new learning are intermixed, this coding obscures a meaningful distinction as these pre-testing studies do not include original learning. That is, the pre-test in such cases does not require episodic retrieval of previous content.

To illustrate, a pre-testing paradigm with new learning provided as feedback involves participants guessing the answer to a question and then being provided the correct answer directly after the guess. Here, feedback is new learning. This can be compared to pre-testing paradigms without feedback where participants guess the answer to all questions, do not receive feedback, and are then provided the new materials. This subsequent block is new learning. Although these are distinctions between the two paradigms, neither contains an original learning phase. In contrast, other studies, such as those of Davis and Chan (2015) and Davis et al. (2017), interleaved original learning and new learning. Participants studied materials, such as a face-name pair, attempted to retrieve part of the material, for example face-?, and then attempted to learn new material, e.g., face-occupation. This distinction is important because research has highlighted that interleaving may carry a negative effect on new learning if participants preferentially encode the original materials over the new materials.

In the Chan et al. (2018) meta-analysis, the interleaved subgroup contained pre-testing studies which used feedback as new learning. In the current analysis, our interleaved subgroup did not. In total, the following 22 studies were removed from the interleaved subgroup: Grimaldi and Karpicke (2012) Experiments 1 and 2, Hays, Kornell, and Bjork (2013) Experiments 1 and 2, Knight, Ball, Brewer, De-

Table 1. Effect size estimate (Hedges' *g*) from reanalyses of Chan et al.'s (2018) dataset with publication bias adjustments, 95% confidence interval in brackets. Non-significant effects are those where 0.00 is included in the confidence interval.

Design	Random effects estimate	Publication bias adjustment 3PSM	Publication bias adjustment PET
Overall	0.44 [0.35, 0.53]	0.02 [-0.14, 0.18]	-0.02 [-0.26, 0.22]
Standard	0.75 [0.66, 0.85]	0.52 [0.34, 0.71]	-0.08 [-0.30, 0.14]
Pre-test	0.34 [0.19, 0.50]*	N/A**	N/A**
Interleaved	-0.56 [-0.73, -0.40]*	-0.55 [-0.75, -0.35]	0.02 [-0.30, 0.33]

*The random effects meta-analysis estimates are different from those reported by Chan et al. due to differences in coding. Their estimate for the pre-test design was 0.35 [0.20, 0.50] and for the interleaved design was -0.02 [-0.16, 0.13].
** Symmetric funnel plot and non-significant Egger's test indicated that publication bias was unlikely in the pre-testing designs.

Witt, and Marsh (2012) Experiments 1a and 1b, Kornell, Hays, and Bjork (2009) Experiments 1-6, Potts and Shanks (2014) Experiments 1-3, Potts (2013) Experiments 4, 6-8, and Vaughn, Hausman, and Kornell (2016) Experiments 1-4. This recoding permitted us to calculate a separate meta-analytic effect size estimate for interleaved designs. As we show below, Chan et al.'s treatment of interleaving as a variable rather than as a category led them to conclude that its overall meta-analytic effect was very close to zero. In contrast, we show that interleaved designs yield a strong negative effect size, though with clear evidence of publication bias.

Within their meta-analysis Chan et al. (2018) noted that the standard, pre-test and interleaved designs produce radically different effects. We argue that rather than pooling these studies, analysis favours subcategorization treating each as a different pool of studies and as three distinct effects. This has the additional advantage of reducing between-study heterogeneity, allowing a more accurate assessment of publication bias. As will be shown, these effects are: (1) a medium-sized TPNL effect assessed using standard designs but with clear evidence of publication bias; (2) a small but reliable pre-testing effect, assessed using a pre-testing design (referred to by Chan et al. as a single-list blocked and interleaved design); and (3) a negative interleaved-testing effect, measured through interleaved designs, again with evidence of publication bias.

Extension of Chen et al.'s analyses

We conducted a replication and extension of the Chan et al.'s (2018) meta-analysis treating the studies using standard, pre-testing, and interleaved designs as distinct, using the same random-effects model and parameters described above. The original and corrected estimates are reported in Table 1. The funnel plots demonstrating the differences between the TPNL, pre-testing, and interleaving effects are depicted in Figure 1, panels B-D. Our analyses reproduced the effect sizes described in the Chan et al. meta-analysis for the standard and pre-testing paradigms. The interleaved sample used in our analysis was different to that used by Chan et al., as were our results.

TPNL in standard designs. When treated as an independent effect, testing robustly potentiated new learning relative to comparison tasks in standard designs, $k = 84$, $t(83) = 15.53$, $p < .0001$, $I^2 = 70.5\%$. Given the large heterogeneity, we conducted an influence analysis and inspected a Baujat plot which did not highlight any significant outliers contributing disproportionately to between-study heterogeneity. Importantly, Egger's test found a very strong relationship between effect size and standard error, $B = 3.25$, $t(83) = 6.70$, $p < .0001$. The 3PSM adjusted estimate was significantly smaller, but still robust and reliable (see Figure 1, panel B), and a likelihood ratio test confirmed that the adjusted model fit the data better than the unadjusted model, $\chi^2(1) = 15.85$, $p < 0.001$. In contrast the PET adjusted estimate as shown by the intercept of the red line was not only much smaller, but was also not significantly different from zero, $t(83) = 0.71$, $p = .48$. These results indicate noteworthy uncertainty surrounding the magnitude and existence of the standard TPNL effect.

Pre-testing designs. Experiments employing pre-testing designs ($k = 45$) only provided a small-to-medium benefit to learning, $t(44) = 4.47$, $p < .0001$, $I^2 = 87.8\%$. However Egger's test was not significant, consistent with a symmetric funnel plot (see Figure 1, panel C) and negligible publication bias: $B = 1.28$, $t(44) = -0.90$, $p = .37$. An influence analysis and inspection of Baujat plot highlighted six outliers disproportionately contributing to the overall effect. When these were excluded from the analysis ($k = 39$), the effect size estimate was larger and heterogeneity was substantially reduced, $g = 0.44$, CI [0.32, 0.55]), $t(38) = 7.50$, $p < .0001$, $I^2 = 73.2\%$. An inspection of those studies, however, reveals that rather than being outliers, they are indicative of a variable that affects the pre-testing effect but not TPNL: The presence of a *lag* between the initial test and new learning completely eliminates the pre-testing effect but has little impact on TPNL. This is discussed later.

Interleaved designs. There were few studies using an interleaved design ($k = 22$). Although the overall effect size suggests that interleaving produces a medium-sized impairment to new learning, $g = -0.56$, CI [-0.73, -0.40]), $t(21) = -7.07$, $p < .0001$, with relatively little between-study heterogeneity ($I^2 = 49.0\%$), Egger's test was significant, $B = -2.42$, $t(21) = -3.46$, $p = .003$, suggesting publication bias (see panel D, Figure 1). The 3PSM adjusted estimate was however similar in magnitude ($g = -0.55$, CI[-0.35, -0.75], $Z = -5.45$, $p = .008$), and the adjusted fit was not significantly better than that of the unadjusted model, $X^2(1) = 0.02$, $p =$

0.90. Once again, however, the PET adjusted estimate was not only significantly smaller in magnitude, but was also not different from zero, (g = 0.02, CI [-0.30, 0.33]), $t(21)$ = 0.124, p = .90). As with the standard design, this analysis of publication bias calls into question the reliability of the uncorrected meta-analytic effect.

Summary. Chan et al. concluded that TPNL, pooled across all designs, is a robust phenomenon, and reached a similar conclusion about studies employing standard and pre-testing designs. They concluded that interleaving designs yielded no overall effect. Our extensions of their analysis, which include corrections for publication bias, point to a very different set of conclusions for at least two of these four cases: (1) There is no overall effect when all designs are pooled; both the 3PSM and PET methods yield corrected estimates that are not significantly different from zero; (2) Experiments employing the standard design fall short of providing convincing evidence of TPNL; although the corrected estimate is significantly greater than zero with one correction method (3PSM), it is not with the other (PET); (3) We concur with Chan et al.'s conclusions regarding pre-testing designs: these yield a small but reliable effect; (4) Finally, we find a medium-sized negative effect for interleaved designs, but this effect must be regarded as only weakly supported by the available evidence: it survives application of one correction method (3PSM) but not the other (PET). It is important to note that publication bias adjustment is complex, and estimates derived from the 3PSM and PET methods should be interpreted cautiously. Although Carter et al.'s (2019) application indicates that both are appropriate adjustment methods, it is notable that they produced dramatically different estimates for the standard and interleaved designs. Rather than treat these estimates as robust metrics of the true effect size, we argue that it is more appropriate to interpret them as markers of significant uncertainty regarding the magnitude and existence of the effects described in the Chan et al. (2018) meta-analysis.

The source of funnel plot asymmetry is also complex. Although usually associated with publication bias (Begg & Berlin, 1988; Light & Pillemer, 1984), there are alternative reasons why a funnel plot may be asymmetric (Egger et al., 1997; Sterne et al., 2000). For example, if the true effects are heterogeneous but researchers engage in pilot work that provides an estimate of the likely effect size, then testing a smaller sample when the anticipated effect is larger would be sound experimental practice. Within the Chan et al. (2018) dataset, there is a small suggestion that pilot work might have informed effect size estimates and sample size planning: in a random selection of 20 studies, 5 mentioned pilot work. But even if this pilot work provided effect size information that informed sample size planning (which is explicitly stated in none of these cases), the majority of studies offer no evidence of sample size planning.

Evaluating Theoretical Accounts

Should effects from standard, pre-testing, and interleaved designs be combined? As previously mentioned, we believe that the standard, pre-testing and interleaved effects recruit partially distinct cognitive mechanisms and that pooling these effects is unwarranted. Chan et al. (2018) did note that the three designs produced significantly different effect sizes, and that the effects were also differentially impacted by moderators. However, they did not adequately question whether this was indicative of a contrast in core mechanisms and grouped the impact of moderators in an overall evaluation of mechanisms for an aggregated "TPNL" effect. This is important as we have demonstrated that the aggregated effect is not reliable when publication bias is accounted for, undermining any discussion on potential mechanisms.

To demonstrate this point, we replicated the theoretical evaluation conducted by Chan et al. (2018), which evaluated the aggregated effect, but applied the analysis separately to the standard, pre-testing and interleaved samples. Full details are provided in their article, but in short, Chan et al. coded each study on the number of characteristics that an account predicts would boost the aggregated TPNL effect, and meta-regressed observed effect sizes on these characteristics. This permits an assessment of how accurately each account predicted differences in effect sizes. For example, Chan et al. argued that integration theories predict that interleaving, using a comparison task that is not restudy, using related materials, using an initial test format that is either episodic retrieval or pre-testing, and higher initial test performance, are all methodological characteristics that should boost TPNL. If a study had five of these characteristics, it would score 5 on the integration theory score. In contrast, if it had none of these characteristics, it would score 0. Integration theories reason that higher scores predict larger TPNL effects as they have more of these positive characteristics. As Chan et al. evaluated four theories of TPNL, four scores were computed, one for each: resource theories, metacognitive theories, context theories, and integration theories. The conclusion drawn from their meta-regression was:

> In sum, according to the qualitative assessment, both resource theories and integration theories received considerable support from our data. Results from the metaregression analysis and dominance analysis largely corroborated this conclusion, but they also established that integration theories provided better predictions for the data than did all other theories, including resource theories. (p. 1133)

This conclusion, we argue, is unwarranted. The numeric data of our reanalysis are presented in Table 2. To summarise, when looking at the pooled effect, we reproduced their results, and the order of the regression coefficients was Integration > Resource > Context > Metacognitive. However, the patterns were very different for the subgroups. In the standard designs, the order is Resource > Integration > Metacognitive = Context. In the pre-testing designs it is Resource > Metacognitive > Integration > Context, and in the interleaved designs the order is Integration > Metacognitive > Resource = Context. The conclusion that both integration and resource theories are better supported by the data is only partially true for the pre-testing dataset, but is not for either standard or interleaved datasets.

We do not draw any major theoretical conclusions about the mechanisms for the different effects from this ordering.

Table 2. Results of the Meta-Regression Analysis

Design and Theory	*B*	95% CI	*t*	*p*-value	R^2
Overall					
Resource theories	0.35	[0.27, 0.43]	8.37	<.0001	0.38
Metacognitive theories	0.15	[0.04, 0.25]	2.77	0.01	0.06
Context theories	0.22	[0.06, 0.38]	2.73	0.01	0.05
Integration theories	0.37	[0.29, 0.45]	8.96	<.0001	0.41
Standard					
Resource theories	0.10	[-0.02, 0.22]	1.72	0.09	0.04
Metacognitive theories	-0.01	[-0.16, 0.14]	-0.13	0.90	0.00
Context theories	-0.01	[-0.16, 0.13]	-0.19	0.85	0.00
Integration theories	0.04	[-0.11, 0.18]	0.51	0.61	0.00
Pre-test					
Resource theories	0.33	[0.11, 0.54]	3.11	0.00	0.19
Metacognitive theories	0.26	[-0.14, 0.66]	1.30	0.20	0.02
Context theories	0.11	[-0.16, 0.38]	0.81	0.42	0.00
Integration theories	0.20	[-0.04, 0.45]	1.69	0.10	0.05
Interleaved					
Resource theories	0.00	[-0.52, 0.52]	0.01	0.99	0.00
Metacognitive theories	-0.07	[-0.27, 0.14]	-0.70	0.49	0.05
Context theories	0.00	[-0.52, 0.52]	-0.01	0.99	0.00
Integration theories	0.13	[-0.35, 0.60]	0.55	0.59	0.00

It is likely the meta-regression for each subgroup is underpowered[1]. However, our analysis demonstrates that any strong theoretical conclusions regarding what mechanisms govern TPNL are unjustified. The support for the different accounts is substantially different across the subgroups. This is particularly evident in the different ordering of the regression weights of the moderator variables for Resource and Integration theories across the standard and interleaved designs.

Concluding Remark

In conducting a meta-analysis, how studies are pooled, assessing whether there is too much heterogeneity for valid aggregated analysis and interpretation, and testing (or correcting) for publication bias are important factors affecting the inferences that can be drawn. On all these counts, we argue that the data from Chan et al.'s (2018) meta-analysis suggest that pooling was inappropriate, resulting in the combination of three very different effects, that the heterogeneity made it impossible to quantify an overall effect, and that there was clear evidence of publication bias that was not adequately addressed. As such, we propose that TPNL is best treated as three distinct effects which are measurable via the impact of previous retrieval practice on new learning in standard designs, the impact of guessing and semantic generation on learning in pre-testing designs, and the impairment of new learning following intermixing original and new learning in interleaved designs.

The results from the current analysis also question the reliability of the TPNL effects in standard and interleaved designs. In both datasets there was evidence of significant publication bias. It would be inappropriate to infer that these effects do not exist, and that is certainly not our conclusion. Correcting for bias in meta-analysis is a complex issue. Although we employed Carter et al.'s (2019) state-of-the-art method for selecting suitable correction methods, the different methods employed in our analyses reached different conclusions about the true underlying effect: for both the standard and interleaved designs, 3PSM suggested that there is a residual true effect after correction for publication bias, but PET did not. One possible solution to this issue is the use of pre-registered replication studies. Pre-registration would permit TPNL to be precisely estimated in the absence of publication bias.

1 As a separate issue, the validity of this meta-regression analysis, compared to more conventional theory-testing approaches (Farrell & Lewandowsky, 2018), is unknown. This method, in which studies are coded in an all-or-none way according to features that a given theory regards as important, has to the best of our knowledge never been evaluated in simulation studies using data generated by different theories, to ascertain whether it permits the true theory to be recovered.

Funding

This research was supported by a grant from the United Kingdom Economic and Social Research Council (ES/S014616/1).

Data Accessibility Statement

R code used to conduct the analyses can be found here: https://osf.io/2s7dt/

Competing Interests

Authors have no conflicts of interest to disclose.

Contributions

Contributed to conception and design: Shaun Boustani, David Shanks

Contributed to analysis and interpretation of data: Shaun Boustani, David Shanks

Drafted and/or revised the article: Shaun Boustani, David Shanks

Approved the submitted version for publication: Shaun Boustani, David Shanks

Submitted: December 31, 2020 PST, Accepted: January 18, 2022 PST

REFERENCES

Begg, C. B., & Berlin, J. A. (1988). Publication bias: A problem in interpreting medical data. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, *151*(3), 419. https://doi.org/10.2307/2982993

Borenstein, M. (2009). *Introduction to meta-analysis*. John Wiley & Sons. https://doi.org/10.1002/9780470743386

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*(2), 115–144. https://doi.org/10.1177/2515245919847196

Chan, J., Meissner, C., & Davis, S. (2018). Retrieval potentiates new learning: A theoretical and meta-analytic review. *Psychological Bulletin*, *144*(11), 1111–1146. https://doi.org/10.1037/bul0000166

Davis, S. D., & Chan, J. C. K. (2015). Studying on borrowed time: How does testing impair new learning? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*(6), 1741–1754. https://doi.org/10.1037/xlm0000126

Davis, S. D., Chan, J. C. K., & Wilford, M. M. (2017). The dark side of interpolated testing: Frequent switching between retrieval and encoding impairs new learning. *Journal of Applied Research in Memory & Cognition*, *6*(4), 434–441. https://doi.org/10.1016/j.jarmac.2017.07.002

Egger, M., Smith, G. D., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *BMJ*, *315*(7109), 629–634. https://doi.org/10.1136/bmj.315.7109.629

Farrell, S., & Lewandowsky, S. (2018). *Computational modeling of cognition and behavior*. Cambridge University Press. https://doi.org/10.1017/CBO9781316272503

Grimaldi, P. J., & Karpicke, J. D. (2012). When and why do retrieval attempts enhance subsequent encoding? *Memory & Cognition*, *40*(4), 505–513. https://doi.org/10.3758/s13421-011-0174-0

Hartung, J., & Knapp, G. (2001). On tests of the overall treatment effect in meta-analysis with normally distributed responses. *Statistics in Medicine*, *20*(12), 1771–1782. https://doi.org/10.1002/sim.791

Hays, M. J., Kornell, N., & Bjork, R. A. (2013). When and why a failed test potentiates the effectiveness of subsequent study. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(1), 290–296. https://doi.org/10.1037/a0028468

Hedges, L. V., & Vevea, J. L. (1998). Fixed- and random-effects models in meta-analysis. *Psychological Methods*, *3*(4), 486–504. https://doi.org/10.1037/1082-989x.3.4.486

Iyengar, S., & Greenhouse, J. (1988). Selection models and the file drawer problem. *Statistical Science*, *3*(1), 109–117. https://doi.org/10.1214/ss/1177013012

Knight, J. B., Ball, B. H., Brewer, G. A., DeWitt, M. R., & Marsh, R. L. (2012). Testing unsuccessfully: A specification of the underlying mechanisms supporting its influence on retention. *Journal of Memory and Language*, *66*(4), 731–746. https://doi.org/10.1016/j.jml.2011.12.008

Kornell, N., Hays, M. J., & Bjork, R. A. (2009). Unsuccessful retrieval attempts enhance subsequent learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(4), 989–998. https://doi.org/10.1037/a0015729

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

McDermott, K. B. (2021). Practicing retrieval facilitates learning. *Annual Review of Psychology*, *72*(1), 609–633. https://doi.org/10.1146/annurev-psych-010419-051019

Paule, R. C., & Mandel, J. (1982). Consensus values and weighting factors. *Journal of Research of the National Bureau of Standards*, *87*(5), 377. https://doi.org/10.6028/jres.087.022

Peters, J., Sutton, A., Jones, D., Abrams, K., Rushton, L., & Moreno, S. (2010). Assessing publication bias in meta-analyses in the presence of between-study heterogeneity. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *173*(3), 575–591. https://doi.org/10.1111/j.1467-985x.2009.00629.x

Potts, R. (2013). *Memory interference and the benefits and costs of testing* [Unpublished dissertation]. University College London, London, UK.

Potts, R., & Shanks, D. R. (2014). The benefit of generating errors during learning. *Journal of Experimental Psychology: General*, *143*(2), 644–667. https://doi.org/10.1037/a0033194

Roediger, H. L., & Karpicke, J. D. (2006). Test-enhanced learning. *Psychological Science*, *17*(3), 249–255. https://doi.org/10.1111/j.1467-9280.2006.01693.x

Sidik, K., & Jonkman, J. (2002). A simple confidence interval for meta-analysis. *Statistics in Medicine*, *21*(21), 3153–3159. https://doi.org/10.1002/sim.1262

Stanley, T. (2008). Meta-regression methods for detecting and estimating empirical effects in the presence of publication selection. *Oxford Bulletin of Economics and Statistics*, *70*(1), 103–127. https://doi.org/10.1111/j.1468-0084.2007.00487.x

Sterne, J. A. C., Gavaghan, D., & Egger, M. (2000). Publication and related bias in meta-analysis: Power of statistical tests and prevalence in the literature. *Journal of Clinical Epidemiology*, *53*(11), 1119–1129. https://doi.org/10.1016/s0895-4356(00)00242-0

Szpunar, K., Khan, N., & Schacter, D. (2013). Interpolated memory tests reduce mind wandering and improve learning of online lectures. *Proceedings of the National Academy of Sciences*, *110*(16), 6313–6317. https://doi.org/10.1073/pnas.1221764110

Terrin, N., Schmid, C., Lau, J., & Olkin, I. (2003). Adjusting for publication bias in the presence of heterogeneity. *Statistics in Medicine*, *22*(13), 2113–2126. https://doi.org/10.1002/sim.1461

Vaughn, K. E., Hausman, H., & Kornell, N. (2016). Retrieval attempts enhance learning regardless of time spent trying to retrieve. *Memory*, *25*(3), 298–316. https://doi.org/10.1080/09658211.2016.1170152

Vevea, J., & Hedges, L. (1995). A general linear model for estimating effect size in the presence of publication bias. *Psychometrika*, *60*(3), 419–435. https://doi.org/10.1007/bf02294384

Vevea, J., & Woods, C. (2005). Publication bias in research synthesis: Sensitivity analysis using a priori weight functions. *Psychological Methods*, *10*(4), 428–443. https://doi.org/10.1037/1082-989x.10.4.428

Yang, C., Luo, L., Vadillo, M. A., Yu, R., & Shanks, D. R. (2021). Testing (quizzing) boosts classroom learning: A systematic and meta-analytic review. *Psychological Bulletin*, *147*(4), 399–435. https://doi.org/10.1037/bul0000309

SUPPLEMENTARY MATERIALS

Peer Review History

Download: https://collabra.scholasticahq.com/article/31996-heterogeneity-and-publication-bias-in-research-on-test-potentiated-new-learning/attachment/80544.docx?auth_token=DkwVDWlGrZGcbmQgUa5H