Transfer of Test-Enhanced Learning:

Meta-Analytic Review and Synthesis

Steven C. Pan

Timothy C. Rickard

University of California, San Diego

Word count (main text and references): 28,430

This manuscript was accepted for publication in *Psychological Bulletin* on February 27, 2018.

Author Note

Steven C. Pan, Department of Psychology, University of California, San Diego; Timothy C. Rickard, Department of Psychology, University of California, San Diego.

Portions of this research were presented at the 58th Annual Meeting of the Psychonomic Society in Vancouver, BC. This research was supported by an American Psychological

Association (APA) Early Graduate Student Researcher Award and a National Science Foundation (NSF) Graduate Research Fellowship to S. C. Pan.

The authors gratefully acknowledge Shana Carpenter, Kit Cho, Sean Kang, Hal Pashler, John Wixted, and anonymous reviewers for helpful suggestions and/or feedback on earlier versions of this manuscript; Pooja Agarwal, Robert Bjork, Elizabeth Bjork, Andrew Butler, Shana Carpenter, Cho Kin Cheng, Kit Cho, Regina Coles, Ed DeLosh, Luke Eglington, Lisa Fazio, Alice Healy, Reshma Gouravajhala, Mark Huff, Sean Kang, Nate Kornell, Mark McDaniel, Kathleen McDermott, Katherine Rawson, Henry Roediger, Christopher Rowland, Thomas Toppino, Randy Tran, Kalif Vaughn, Peter Verkoeijen, and others for providing data, manuscripts, and/or assistance with the literature search; Andrew Butler, Donald Foss, Scott Hinze, Mark McDaniel, Katherine Rawson, and Jennifer Wiley for helpful discussions; Ikjot Thind for assistance with intercoder reliability; Natalia Cameroni-Adams, Dominic D'Andrea, Charles Dupont, Kellie King, Qiyang Lin, Joshua Lozano, Dania Pagarkar, and Cullin McLean Taggard for assistance with data extraction and/or verification; and to the many others who commented on this work.

Correspondence concerning this article should be addressed to Steven C. Pan, Department of Psychology, University of California, San Diego, La Jolla CA 92093-0109. E-mail: stevencpan@ucsd.edu.

Abstract

Attempting recall of information from memory, as occurs when taking a practice test, is one of the most potent training techniques known to learning science. However, does testing yield learning that transfers to different contexts? In the present article, we report the findings of the first comprehensive meta-analytic review into that question. Our review encompassed 192 transfer effect sizes extracted from 122 experiments and 67 published and unpublished articles (N = 10,396) comprising over 40 years of research. A random-effects model revealed that testing can yield transferrable learning as measured relative to a non-testing reexposure control condition (d = 0.40, 95% CI [0.31, 0.50]). That transfer of learning is greatest across test formats, to application and inference questions, to problems involving medical diagnoses, and to mediator and related word cues; it is weakest to rearranged stimulus-response items, to untested materials seen during initial study, and to problems involving worked examples. Moderator analyses further indicated that response congruency and elaborated retrieval practice, as well as initial test performance, strongly influence the likelihood of positive transfer. In two assessments for publication bias (using PET-PEESE and various selection methods), the moderator effect sizes were minimally affected. However, the intercept predictions were substantially reduced, often indicating no positive transfer when none of the aforementioned moderators are present. Overall, our results motivate a three-factor framework for transfer of test-enhanced learning and have practical implications for the effective use of practice testing in educational and other training contexts.

Keywords: retrieval practice, testing effect, test-enhanced learning, transfer, meta-analysis

Public Significance Statement

The present meta-analysis found that practice testing can result in learning that generalizes to different situations and different test types. That transfer of learning is greatest across test formats, to application and inference questions, to problems involving medical diagnoses, and to tests involving mediator or related word cues. It is weakest to rearranged cues and responses, to unpracticed information that was seen during prior study, and to problems involving worked examples.

Transfer of Test-Enhanced Learning: Meta-Analytic Review and Synthesis

The act of attempting to recall information from memory, as occurs when taking a test, provides not only an assessment of prior learning but also a potent new learning opportunity. That finding is the chief result of more than 200 studies from over a century of research (beginning with Abbott, 1909), in confirmation of earlier anecdotal observations (e.g., James, 1890). Studies showing the benefit of testing for memory – more formally known as *test-enhanced learning*, the *testing effect,* or the *retrieval practice effect* – commonly utilize a three-phase experimental paradigm that begins with (a) initial study of a set of to-be-learned materials (e.g., word lists or text passages), followed by (b) training on those materials via testing or, for comparison purposes, a non-testing reexposure control condition (e.g., restudy), and ending with (c) a final test. On that final test, materials that were initially tested are usually better remembered than those that were not. Test-enhanced learning has been demonstrated across a wide range of materials (for a listing, see Rawson & Dunlosky, 2011; for reviews, see Bjork, 1975; Dempster, 1996; Rickard & Pan, 2017; Roediger & Butler, 2011; Roediger & Karpicke, 2006; Roediger, Putnam, & Smith, 2011; for meta-analyses, see Adescope, Trevisan, & Sundararajan, 2017; Rowland, 2014), with a variety of test types (e.g., McDaniel, Wildman & Anderson, 2012; Pan, Gopal, & Rickard, 2015), with and without correct answer feedback (i.e., being shown the correct answer) after attempting retrieval (e.g., McDaniel, Bugg, Liu, & Brick, 2015; Rowland & DeLosh, 2015b), across a variety of retention intervals (e.g., Carpenter, Pashler, Wixted, & Vul, 2008; McDaniel, Howard, & Einstein, 2009), and with individuals of diverse ages (e.g., Carpenter et al., 2016; Jones et al., 2015; Meyer & Logan, 2013) and different memory abilities (e.g., Agarwal, Finley, Rose, & Roediger, 2016; Pan, Pashler, Potter, & Rickard, 2015).

Given the strong evidence for its memorial benefits, many cognitive and educational

psychologists now classify testing as among the most effective educational techniques discovered to date. These researchers emphasize that tests are beneficial not just for assessment, but also as powerful learning tools in and of themselves (i.e., in the form of practice or no-stakes tests; for discussions see Benjamin & Pashler, 2015; Bourne & Healy, 2013; Brown, Roediger, & McDaniel, 2014; Fiorella & Mayer, 2015; Karpicke, 2012; McDaniel, Roediger, & McDermott, 2007; Pashler, Rohrer, Cepeda, & Carpenter, 2007; Rawson & Dunlosky, 2012; Roediger & Pyc, 2012). Accordingly, test-enhanced learning is prominently featured in reports on evidence-based training methods from the U.S. National Center for Education and the National Research Council (Druckman & Bjork, 1994; Pashler, Bain, et al., 2007), is highlighted in a recent comprehensive review of effective learning techniques from cognitive and educational psychology research (Dunlosky, Rawson, Marsh, Nathan, & Willingham, 2013), and has begun to attract attention from the mainstream media (e.g., Carey, 2013; Lahey, 2014; Paul, 2016).

Most studies of test-enhanced learning feature identical materials (i.e., test questions) during both initial and final tests. Although important theoretically and in some educational contexts (e.g., simple arithmetic or vocabulary; memorization of critical facts, equations, or reactions in multiple STEM fields), in numerous situations more flexibly applicable learning is needed. For instance, instructors frequently eschew revealing exam questions during classroom lessons, practice quizzes, and other training activities (e.g., Balch, 1998; Popham, 2011; Wooldridge, Bugg, McDaniel, & Liu, 2014; Mayer, 2009). Moreover, in ecologically valid circumstances one cannot expect always having to recall the same information or correctly anticipating the manner in which previously learned information will need to be used. Instead, different information is presented or needs to be retrieved, previously learned and new information must be integrated, or prior learning must be synthesized in order to arrive at a correct answer or solution. Further, in a variety of circumstances it can be impractical to test on

all to-be-learned materials; as such, it would be desirable to know if and when the benefits of testing on a subset of materials can yield benefits for other, not directly tested materials.

The aforementioned scenarios raise the issue of *transfer of learning* (i.e., the use of prior learning in a new context or contexts; for a more detailed definition and specific examples, see the next section). Transfer of learning is commonly described as a paramount goal or even the "holy grail" of education (Druckman & Bjork, 1994; Haskell, 2001; McDaniel, 2007). The critical question arises: beyond aiding retention, does testing enhance the transfer of learning (henceforth, *transfer*) to new contexts?

Two reviews of the test-enhanced learning literature, Roediger and Butler (2011) and Roediger, Putnam, et al. (2011), included subsections on transfer; in both articles, the conclusion (based on the limited evidence available) was that testing does indeed generate transferrable learning. Roediger, Putnam et al. included transfer among their "ten benefits of testing", of which two were "testing improves transfer of knowledge to new contexts" and "testing can facilitate retrieval of material that was not tested" (pp. 14-20). Similar statements can be found in recent empirical articles on test-enhanced learning, including Butler (2010), Rohrer, Taylor, and Sholar (2010), and Carpenter and Kelly (2012), as well as in articles on the technique written for the general public (e.g., Lahey, 2014; Paul, 2016; Swaminathan, 2006).

Carpenter (2012), in a brief review that was the first and, prior to this writing, only paper to specifically focus on this topic, highlighted over two dozen studies and concluded that testing can yield transferrable learning, but noted that further research is needed to gain a more comprehensive understanding of that transfer. Since that review, the literature on transfer of test-enhanced learning has grown exponentially, now exceeding 70 studies. It contains a diverse set of experiments that vary in terms of transfer contexts (e.g., involving the same vs. different cues; contexts to be further detailed later in this article), types of initial tests (e.g., free vs. cued recall),

and other potentially critical experimental design features (e.g., brief or long retention intervals; between- vs. within-subjects designs, classroom vs. laboratory settings, etc.).

In light of that growth, it is broadly agreed in the field that a new, comprehensive review is needed. In the present article, we address that need through meta-analysis of 192 effect sizes from 122 experiments and 67 articles in which transfer was measured relative to a non-testing reexposure control condition. That analysis provides, for the first time at the level of the literature, statistically-based insight into the conditions under which transfer occurs, important moderating factors, generalizability, and candidate theories.

Definition of Transfer and Relevant Terminology

Drawing on prior literature (e.g., Carpenter, 2012; Gick & Holyoak, 1987; Haskell, 2001; McGeoch, 1942; Roediger, 2007), the definition of transfer used throughout this review is the productive use of prior learning in a novel context. What exactly constitutes a "novel context"? In transfer research, a novel context can potentially refer to any situation that is different in some way from that in which original learning took place (McDaniel, 2007). This may include a different topic, a different goal, a different test type, or any number of other contextual changes (for a taxonomy, see Barnett & Ceci, 2002). For example, if information that is trained via a free recall test is later assessed on a final multiple-choice test, then that final test constitutes a novel context (i.e., transfer across test formats). Alternatively, if prior learning needs to be integrated with new information on a final application test, then that application test constitutes a novel context (i.e., transfer to application questions). In another example, if learners are trained to recall words given specific cues (e.g., given *mother*, recall *child*), and then have to recall those words in response to different cues on a final test (e.g., given *father,* recall *child*), then that final test also constitutes a novel context (i.e., transfer to mediator word cues). The list of novel contexts that potentially involve transfer is limitless.

Some contextual changes are more extensive than others. For instance, a change in test format is typically regarded as less substantial than the combination of a change in subject matter and a switch to application questions. In the transfer literature, the range of possible novel contexts is often dichotomized into *near transfer* (i.e., relatively minor) and *far transfer* (i.e., extensive or multiple changes) categories (Barnett & Ceci, 2002; Perkins & Salomon, 1994). Some transfer researchers argue that relatively minor contextual changes (i.e., "near" transfer) constitute "ordinary learning" and should not be considered as involving transfer (Perkins & Salomon, 1994), although there is no absolute dividing line between ordinary learning and transfer. Drawing on that precedent, in this review we did not consider studies in which the contextual change was solely the passage of time or a change in physical location as involving transfer.[1] For the current purposes, such changes were too minor to constitute meaningful transfer (i.e., they represent ordinary learning). Overall, our review encompassed a wide range of educationally, practically, and theoretically meaningful transfer contexts – including six major transfer categories that span from "near" to "far" transfer (namely transfer *across test formats*, to *stimulus-response rearrangement*, to *untested materials seen during initial study*, to *application and inference questions*, of *problem-solving skills*, and to *mediator and related word cues*; each are defined in subsequent sections of this review) – that comprise the vast majority of the literature on transfer of test-enhanced learning to date.

The test-enhanced learning paradigm. Studies in the test-enhanced learning literature

---

[1] For studies involving transfer of test-enhanced learning, the retention interval between the training phase and final test is typically equivalent across the following categories: (a) items that were not tested (e.g., the restudied items) during training and only tested on the final test, (b) items that were tested during training and tested in an identical way on the final test (yielding the testing effect as defined in this review), and (c) items that were tested during training and then tested in a different context on the final test (yielding the transfer effect). Hence, the effect of retention interval on final test performance should be similar for the non-testing reexposure control and transfer conditions on the final test, the two conditions through which the transfer effect is measured.

commonly feature a three-phase experimental paradigm. This paradigm is described as follows. First, after (a) an initial study phase on a set of to-be-learned materials, which we will refer to as *initial study*, those materials are (b) practiced in a *training phase* via testing or a non-testing method. We will use *initial test* to describe training through testing, and the non-testing method will be described generally as the *non-testing reexposure control* (when discussing individual studies, we will refer to the non-testing reexposure control by the task that is used, such as restudy or rereading). Finally, after a common retention interval, prior learning is assessed via (c) a *final test* (i.e., criterial test). That final test allows comparison of learning and retention that occurred via testing vs. the non-testing reexposure control condition. In some cases, the final test includes both transfer and non-transfer questions; when discussing such cases, we will differentiate final test questions or tests that directly assess transfer by using the term *transfer test* (i.e., a final test that specifically focuses on transfer).

Test-enhanced learning vs. transfer of test-enhanced learning. In this review we will also distinguish between the effects of testing where transfer is and is not involved (i.e., testing's effects on transfer vs. verbatim retention). We investigated the former case (i.e., transfer of test-enhanced learning) using quantitative meta-analysis; in a supplementary analysis we compared both types of effects (test-enhanced learning vs. transfer of test-enhanced learning). For brevity, the term *testing effect* will be used to refer to the case of identical contexts, retrieval cues, and required responses on initial and final tests (which could also be described as "conventional test-enhanced learning" or a "retention effect"),[2] and the term *transfer effect* will apply to the case of differences in either cues or required responses (or both) on the initial and final tests (i.e., the

[2] Both testing and transfer effects can be assumed to be a result of testing's effects on either learning, memory, or both. Thus, "test-enhanced learning" and "transfer of test-enhanced learning", respectively, would be perhaps the most accurate descriptors (and would have been used in this review if not for their length).

effects of testing where transfer is involved; a synonym would be "transfer of test-enhanced learning"). In this review, both testing and transfer effects are measured relative to final test performance in a non-testing reexposure control condition. A *positive transfer effect* (or, as shorthand, simply a *transfer effect*) will refer to final test performance that is superior to that in the control condition, and a *negative transfer effect* will refer to the opposite case (McGeoch, 1942; Haskell, 2001). The use of "transfer" as a verb can be assumed to refer to a statistically positive transfer effect.

Theorizing Relevant to Transfer of Test-Enhanced Learning

A comprehensive discussion of all theories and research perspectives from the test-enhanced learning and broader transfer literatures is beyond the scope of this review (for discussions of the former, see Delaney, Verkoeijen, & Spirgel, 2010; Karpicke, Lehman, & Aue, 2014; Roediger & Butler, 2011; Roediger & Karpicke, 2006; van den Broek et al., 2016; for coverage of the latter, see Cormier & Hagman, 1987; Ellis, 1965; Haskell, 2001; McGeoch, 1942; Mestre, 2005; Singley & Anderson, 1989). However, several theoretical perspectives provide relevant background and are briefly summarized here.

Perspectives from the test-enhanced learning literature. Although many theoretical accounts of test-enhanced learning do not directly address transfer (e.g., Halamish & Bjork, 2011; Kornell, Bjork, & Garcia, 2011; Mozer, Howe, & Pashler, 2004), the following three theories suggest a process mechanism that incorporates it. First, the *elaborative retrieval hypothesis* (Carpenter & DeLosh, 2006; see also Carpenter, 2009) posits that a process of spreading activation occurs during the search for correct answers on tests (cf. ACT-R, Adaptive Control of Thought-Rational and SAM, Search of Associative Memory; Anderson, 1996; Collins & Loftus, 1975; Raajmakers & Shiffrin, 1981); as a result, multiple retrieval routes are created which aid later recall, resulting in the testing effect. Transfer effects may also result from the

same mechanism: when information that is semantically related to previously tested information needs to be recalled on a transfer test, the process of spreading activation that presumably occurred during initial testing increases the likelihood that such information will be recallable as well (Carpenter, 2011; Chan, 2009; Chan, McDermott, & Roediger, 2006; Cranney, Ahn, McKinnon, Morris, & Watts, 2009). Second, the *mediator effectiveness hypothesis* (Pyc & Rawson, 2009), posits that *mediators* (i.e., a word, phrase, or concept that links a cue with a target) activated during testing support improved final test performance. By that account, testing can also be expected to improve performance when the mediators themselves, or other information linked via mediators, need to be recalled on a transfer test (Coppens, Verkoeijen, Bouwmeester, & Camp, 2016). Finally, the recently-proposed *dual memory* theory of test-enhanced learning (Rickard & Pan, 2017) constitutes a viable framework from within which to account for results in some cases. According to that theory, test-enhanced learning stems from the fact that two routes to retrieval are accessible for a tested response (i.e., via "study memory" from initial study or "test memory" from the initial test). However, when different responses are required on a transfer test, that theory, in a slightly elaborated form (see Rickard & Pan, 2018), predicts that only study memory is accessible. Under such circumstances, testing is predicted to yield no positive transfer relative to a non-testing reexposure control.

Besides those process-based accounts, Roediger, Putnam, et al. (2011); McDaniel, Thomas, Agarwal, McDermott, and Roediger (2013); Avci (2011; see also Schmidt & Bjork, 1992); and others have speculated that testing may generally yield learning that is more "flexible", improves overall understanding, and/or increases higher-order processing. These descriptive accounts imply that test-enhanced learning will generally yield transfer. Additionally, McDaniel, Howard, et al. (2009); McDaniel and Little (in press); Nguyen and McDaniel (2016); Pan, Gopal, et al. (2015); Pan and Rickard (2017); and van Eersel, Verkoeijen,

Povilenaite, and Rikers (2016) have suggested that activities associated with but separate from the act of testing itself (such as the processing of feedback, more effective subsequent restudy, and more focused attention) may influence the likelihood of transfer of test-enhanced learning.

Perspectives from the broader transfer literature. Transfer of test-enhanced learning intersects with a long-running debate in the broader transfer literature between two prominent and highly influential theoretical perspectives: (a) the *identical elements* and related similarity-based models of transfer, and (b) the *general principle* and other abstractionist models. Those perspectives make contrasting predictions as to the prevalence of transfer (for related discussions, see Allport, 1937; Barnett & Ceci, 2002; Detterman, 1993; Dudai, 2007; Healy, 2007; Kelly, 1967; Mestre, 2005; Sternberg, 1993). In the former, transfer is commonly restricted to situations in which the training and transfer contexts are highly similar to one another (Thorndike, 1906; see also Ebbinghaus, 1885). That similarity may be at the level of cues, responses, available knowledge, mental states, and/or abstract mental representations (for discussions see Morris, Bransford, & Franks, 1977; Rickard & Bourne, 1996; Rickard, Healy & Bourne, 1994; Thorndike, 1906; Tulving, 1970, 1984; Singley & Anderson, 1989). In contrast, the general principle and other abstractionist models suggest that the learning of underlying principles (e.g., properties of actions, operations, perceptions, etc.) can facilitate transfer to contexts that are substantially dissimilar from those that were encountered during training (Judd, 1908; see also Gick & Holyoak, 1980; Simon & Hayes, 1977). According to this perspective, transfer can be increased by making learners aware of relevant information needed for successful transfer (e.g., by training with multiple or varied examples, or by informing learners to apply relevant information), and especially if it involves common information or an underlying principle.

To accommodate both theoretical perspectives, some transfer researchers have proposed

integrative frameworks. Perkins and Salomon (1994; see also Salomon & Perkins, 1989) proposed that transfer can occur in "low" circumstances when the stimuli are the same or similar to those that were previously learned as well as in "high" circumstances where learning (i.e., a search for general principles) occurs at a more abstract level. Barnett and Ceci (2002) proposed that all transfer, whether through identical elements or general principles, requires successful (a) recognition of the need to transfer prior learning to the new context, (b) recall of the relevant knowledge, and (c) execution of prior learning in the new context. Both integrative frameworks allow for the possibility that successful transfer can be very difficult to obtain in various circumstances.

Method

Literature Search

To obtain a comprehensive list of empirical research studies addressing the transfer of test-enhanced learning, we first conducted a preliminary analysis of recent empirical and review articles, and then undertook an extensive formal literature search. Included were online database searches for peer-reviewed research articles, dissertations, and theses; ancestral searches of empirical and review article reference lists; and listserv queries and correspondence with authors to obtain additional data and unpublished manuscripts. No date restriction was applied during the literature search, which concluded on September 12, 2016.

Preliminary searches. Due to the lack of standard terminology for transfer studies in this literature (initial database searches with the keyword *transfer* in conjunction with *test-enhanced learning* and its synonyms yielded only a portion of the studies that are known to exist), we examined the Carpenter (2012) review article, three reviews of test-enhanced learning with subsections that addressed transfer (Roediger & Butler, 2011; Roediger & Karpicke, 2006; Roediger, Putnam, et al., 2011), as well as recent empirical articles to identify types of studies

that involve testing and transfer but do not explicitly use transfer terminology. That preliminary search revealed that the vast majority of studies involving transfer of test-enhanced learning do not necessarily discuss transfer *per se* (cf. Adescope et al., 2017). Rather, many studies use terms that are specific to the transfer context or contexts under investigation (e.g., test formats). Accordingly, we compiled a list of the different transfer types that have been investigated in the literature to date. Keywords addressing these (for a complete listing, see the next section) were incorporated into the formal database searches.

Database searches. Two online databases were queried for empirical research articles: PsycINFO and ProQuest Dissertations and Theses. Using these databases, a total of 72 separate searches were performed using the keywords *test-enhanced learning, testing effect, practice testing,* and *retrieval practice* in combination with the terms *transfer*, *format, related, application, inference, problem solving, category, classification,* or *visuospatial*. These searches were intended to broadly capture any studies that may involve testing and transfer, plus address studies from the aforementioned major transfer categories. The searches yielded 383 hits; 103 were duplicates, leaving 280 database records (212 peer-reviewed articles and 66 dissertations, dating from as early as 1963 and as recent as 2016) for further examination. These records were entered into a three-stage review process to determine suitability for inclusion in the meta-analyses. That process, based on that detailed in a prior meta-analytic review on an unrelated topic (Pan & Rickard, 2015) and summarized in PRISMA (2009; see also APA Publications and Communications Board Working Group on Journal Article Reporting Standards, 2008) diagram format in Figure 1, was first completed for the PsycINFO database search results and is detailed as follows.

(Figure 1 around here)

The first stage, title-level review, involved both authors of this review separately

screening each title for (a) any mention of test-enhanced learning research, as well as (b) any mention of transfer. If either condition held or if the title was ambiguous, it was flagged for potential inclusion. If the title clearly indicated that the article did not address testing or transfer, or stated that it was a review, commentary, or did not involve the standard test-enhanced learning paradigm (e.g., generation effects, hypercorrection effects, hypermnesia, and retrieval-induced forgetting), it was eliminated from consideration (cf. Rowland, 2014). All articles flagged by at least one rater were retained for the next stage. Of the 212 peer-reviewed articles entered into the first stage of review, 110 were excluded and 102 survived. Overall interrater agreement was good (Cohen's $\kappa = 0.81$).

The second stage, abstract-level review, involved the same two raters separately reading each article abstract to verify whether both conditions (a) and (b) from the first stage applied. Similar to the first stage, if either rater determined that the necessary conditions applied or that the abstract was too ambiguous for a definitive rating, then the article was flagged for potential inclusion. Additionally, if the abstract indicated that only clinical populations were involved, then the article was excluded. Of the 102 articles entered into the second stage of review, 41 were excluded and 69 survived. Overall interrater agreement was $\kappa = 0.83$.

The third and final stage, article-level review, involved the first author of this review examining the full text of each article to determine whether it unambiguously met a set of five inclusion criteria (which are detailed later in this section) to qualify for meta-analysis, as well as to verify that it did not violate any of the exclusion rules from the preceding stages. In nine instances where an article contained ambiguities, the final inclusion decision was made by both authors discussing and arriving at mutual agreement. Of the 69 articles entered into the third stage, 35 were excluded and 34 survived.

Owing to the good interrater agreement that was observed for the PsycINFO search

results, the 66 unpublished results from the ProQuest Dissertations and Theses database were screened by the first author only. Seventeen records survived title- and abstract-level review; of these, six dissertations passed article-level review and were included in the meta-analyses.

Ancestral searches. In an effort to obtain further studies, the reference lists of all studies that survived the three-stage screening process, as well as those of six review articles or chapters addressing test-enhanced learning and/or transfer (including the four aforementioned articles that were consulted in the preliminary searches, as well as the reference lists of articles in Rawson & Dunlosky, 2011 and Rowland, 2014) were examined. Sixty-three unique references were identified in this manner. All of these references survived title- and abstract-level review; 21 survived article-level review and were included in the meta-analyses.

Unpublished studies. To address publication bias and the "file drawer" issue (Strube & Hartmann, 1983), we contacted 52 researchers to request any unpublished studies involving transfer of test-enhanced learning. The list of contacts was drawn from listservs of researchers in the fields of learning, memory, cognition, and instruction, as well as lists of authors of studies already included in the meta-analyses. In response to our request (issued on May 15, 2016), we received 15 responses and obtained the full text of 10 unpublished manuscripts (and were also referred to articles and dissertations that we had already obtained); of these, six met article-level inclusion criteria and were included in the meta-analyses. In four cases (Cho et al., 2017; Eglington & Kang, 2016; additionally, Pan, Hutter, et al., 2018; Rickard & Pan, 2018), we obtained or had an in-progress or partially redacted manuscript; each of these had sufficient information to determine study eligibility and to extract effect size and other necessary data.

Inclusion criteria for the article-level review stage. At the final review stage, all studies from the database and ancestral searches, as well as unpublished works solicited via author correspondence, were screened against a set of five inclusion criteria. The purpose of

these criteria was to verify that all included studies, experiments, or conditions had specific, clearly identifiable experimental design features and contained sufficient data for quantitative meta-analyses. Exclusion of individual studies or experiments was done solely on the basis of these criteria and was not the result of any assessment of study quality or outcome. The five criteria were:

1. *The most common three-phase test-enhanced learning paradigm must have been used.* This paradigm, which we noted earlier, involves three phases: first, initial study of to-be-learned materials; second, an intervening training phase on those materials which features a testing vs. a non-testing reexposure control manipulation; and third, a final test. This criterion excluded studies which featured unambiguously different sequences of events, dropout schedules, or had the presentation of new and different to-be-learned information during the training phase (e.g., studies of test-potentiated new learning), as well as studies of adjunct prequestions (for reviews of that literature, see Anderson & Biddle, 1975; Frase, 1968; Hamaker, 1986).
2. *Transfer must have been assessed relative to a non-testing reexposure control.* Multiple types of non-testing reexposure controls have been used in the test-enhanced learning literature, including restudy (or rereading), concept mapping, highlighting, and notetaking (among those, restudy is the most common). The requirement that a non-testing reexposure control be used reflected prior assertions (e.g., Carrier & Pashler, 1992; Carpenter & DeLosh, 2006; Kuo & Hirshman, 1996; Rowland, 2014) that studies in which testing is compared against a no-training condition (i.e., materials in the control condition were not presented in any form during the training phase) preclude any objective assessment of testing's benefits relative to any other learning strategy (for similar observations on the importance of the control condition in the broader transfer

literature, see McGeoch, 1942). From an educational standpoint, it is more meaningful to examine whether testing can yield transfer relative to a non-testing learning activity rather than no learning activity. Included studies fell into one of two widely-used experimental design types. In the first type, only the transfer effect is assessed on the final test. In the second type, both testing and transfer effects are assessed on the final test.

3. *Transfer must have been specifically assessed on the final test and separately reported*. Performance on final test questions that address transfer must have been reported apart from any final test questions that did not involve transfer. Studies in which data from transfer and non-transfer questions were not separated were excluded on this basis (for related discussion see Butler, 2010). Additionally, the exact transfer category under investigation (e.g., application questions or stimulus-response rearrangement) must have been clearly identifiable or inferable from the article text.
4. *Proportion correct must have been the dependent measure on the final test.* In the vast majority of studies in the literature, final test performance is reported in terms of proportion correct ranging from 0 to 1.0. Studies that reported data in that manner, as well as studies in which that data could be derived (e.g., number of points earned out of a maximum possible total) were included. For five studies in which a recognition final test was used, proportion correct was used where it was reported as the dependent measure (e.g.., Bies-Hernandez, 2013; Huff, Balota, & Hutchison, 2016; Verkoeijen et al., 2012) or was derivable from reported mean rates of hits (e.g., Carpenter, 2011) or hits minus false alarms (e.g., Jacoby et al., 2010). (It should be noted, however, that proportion correct in the case of recognition does not account for response criterion effects and is an incomplete measure of performance; for discussion see Stanislaw & Todorov, 1999).

5. *All necessary information for effect size calculations must have been reported or derivable.* Effect size, sampling variability, sample size, type of experimental design (between- or within-subjects), and the relevant test statistics and degrees of freedom for pairwise comparisons (e.g., transfer performance in the testing vs. non-testing reexposure control conditions) must have been provided in the text, be derivable from figures in the article (using the pixel-based graphical measurement technique described in Pan & Rickard, 2015), or provided by the authors in response to electronic correspondence.

Non-independent effect sizes within experiments. Experiments within several included studies involved data that was non-independent in some fashion (i.e., multiple transfer conditions compared against the same reference condition, a transfer condition compared against multiple reference conditions, repeated final tests, or data collapsed across conditions). Our criteria for addressing those cases were as follows.

1. *Each transfer effect size must have been derived from non-overlapping experimental means.* In some experiments, there were an uneven number of testing and non-testing reexposure control conditions. These fell into two broad categories: (a) a greater number of testing conditions than non-testing reexposure controls (e.g., a free recall test, a cued recall test, and a restudy training condition), or (b) multiple non-testing reexposure controls compared against a comparatively smaller number of testing conditions (e.g., notetaking and rereading conditions compared against a single test condition). In both circumstances, the multiple pairwise comparisons that are calculable between testing and non-testing conditions are non-independent. For cases involving (a), one pairwise comparison was chosen at random for inclusion in the quantitative meta-analyses (effect sizes that were not included in those analyses are indicated by asterisks in Table 1). For cases involving (b), the non-testing reexposure control condition that most closely

matched restudy (i.e., the most common reference condition in this literature) was included. Where there were multiple reexposure controls involving restudy, the reexposure control condition that was subject to comparable experimental conditions as the transfer condition was included. For example, in Butler (2010; Experiment 2), there were three training conditions: testing, restudy of isolated sentences, and restudy of passages. Given that the testing condition involved viewing feedback in the form of isolated sentences (and not whole passages), the included reexposure control condition involved isolated sentences.

2. *Data from studies with multiple identical final tests must not have been confounded by the effects of a prior identical final test.* In some studies, subjects completed the same exact final test multiple times, such as immediately after training and then again after a delay (e.g., the previously tested items condition in McDaniel, Howard, et al., 2009) or across multiple test blocks. In such situations, only data from the first test for an item was included, as the results of that test constitute the purest measures of the retention and transferability of learning from the training phase. For studies which had independent between-subjects assignment to an immediate and delayed final test, data from both tests were included. For studies which used within-subjects assignment to immediate and delayed final tests, but in which independent and randomly assigned materials (e.g., two separate text passages) were used on the two tests, data from both tests were included (yielding two effect sizes in the analysis dataset).
3. *Data collapsed across conditions or experiments were included and identified as such if no other inclusion criteria were violated.* In some studies, results were only reported for data collapsed across experiments or across conditions (e.g., different retention intervals). Provided that no other inclusion criteria were violated, those results were included as

such in the meta-analyses and noted in Table 1 in the following manner: where multiple experimental conditions were collapsed together, those conditions are denoted with a dual cross symbol; where multiple experiments were collapsed together, the experiment numbers are presented side-by-side in the table; where multiple retention intervals are collapsed together, the delay interval in hrs. is the average of those intervals.

Further criteria for studies of transfer across test formats. Two additional rules applied to studies of transfer across test formats: test format must have been the only change between the initial and final test (and not a change in assessed content). Studies excluded on this basis remained eligible for inclusion in other categories (throughout the dataset, each effect size was included in only one category). Additionally, studies in which subjects completed a final test in the same format as during training (i.e., a test condition), plus completed another final test in a different format as during training (i.e., a transfer condition), were not eligible for inclusion if the test condition preceded the transfer condition. This rule was implemented to avoid including any data in which the effect of a change in final test format was contaminated by a preceding final test in which there was no change in format.

Outliers. We did not specifically identify, nor exclude, outlier effect sizes. All data that qualified according to the aforementioned inclusion criteria were analyzed.

Missing or incomplete information. We contacted 13 authors to request clarifications or additional data; all but one responded, and nine authors were able to provide the requested information within the requested three-month period. In one other case (Coppens et al., 2016), we were able to derive the necessary information from a dataset made publicly available on the Open Science Framework.

Summary of literature search results. Overall, 67 studies comprising 192 transfer effect sizes from 122 experiments met the criteria for inclusion in the overall and category-level

meta-analyses. Of these, 53 studies had been published and 14 were unpublished by the conclusion of the literature search period. The publication, completion, or submission dates of these studies ranged from 1975 to 2016, with the vast majority (60 studies) finished in 2006 or later. All studies but one (Zhou, Ma, Li, & Cui, 2013) were written in English. Nearly all were performed using samples recruited from young adult (i.e., university student) populations; exceptions included 3 studies involving elementary school children, 2 studies with high school students, and 2 studies with older (50-66 yrs in age) adults. Descriptive and statistical information for each study, including stimulus type, delay interval, test format, condition identifiers, sample size, effect size, and sampling variability, are included in Table 1. Forest plots depicting each effect size across the reviewed literature are included in Figure 2.

(Table 1 and Figure 2 possibly around here)

Categorizing Studies in the Transfer of Test-Enhanced Learning Literature

The current body of research on transfer of test-enhanced learning can be organized into six major categories. We defined these categories based on piecewise discussions in the current literature and our judgment. Although they should be treated as preliminary, we believe that they reflect the structure of the literature – as well as some of the major distinctions in underlying cognitive processes – to a first approximation. The six categories are presented in the same order throughout this review; this order follows a general pattern of increasing divergence between the initial and final tests (ranging from relatively "near" to, in some circumstances, "far" on the near vs. far transfer dichotomy). One exception is transfer to mediator and related word cues; that category is included last due to its having the fewest articles, which precluded all but the simplest meta-analyses. As is evident below, the categories investigated to date represent a considerable range of contextual changes and involve different types of information being transferred. The categories were defined as follows.

Test format. In this category, the final test format is different from the initial test format, but no other major types of transfer are involved. An example is Kang, McDermott, and Roediger (2007), which included conditions in which subjects trained on previously read text passages via multiple-choice tests (e.g., *"Source confusion is…?* with four answer choice options) and then took final cued recall tests on the same information (e.g., *"Source confusion is…?"* without any provided answer choices); the correct answer (e.g., "*misattributing content of a memory to the wrong source"*) was the same on both tests. Studies in this category may potentially use any of the following four initial or final test formats: *free recall* (i.e., recall as much of a text as one can remember), *cued recall* (i.e., fill-in-the-blank, fragment completion, or short answer questions), *multiple-choice* (with between four to six answer options), and *recognition* (i.e., two-alternative forced choice or scale judgment old/new questions). In the literature, six combinations of transfer across test formats have been investigated: *free recall to cued recall* (e.g., Karpicke & Blunt, 2011), *free recall to recognition* (e.g., Verkoeijen et al., 2012), *cued recall to free recall* (e.g., Halamish & Bjork, 2011), *cued recall to recognition* (e.g., Carpenter, 2011), *cued recall to multiple-choice* (e.g., Nungester et al., 1982), and *multiple-choice to cued recall* (e.g., Pan, Gopal, et al., 2015). Many of these format combinations are further discussed in Duchastel (1981); Foos and Fisher (1991); Hanawalt and Tarr (1961); Hogan and Kintsch (1971); Mandler and Rabinowitz (1981); McDermott, Agarwal, D'Antonio, Roediger, and McDaniel (2014); Rickard and Pan (2017); Runquist (1983); Smith and Karpicke (2014); and Wenger, Thompson, and Bartling (1980).

Stimulus-response rearrangement. In this category, all of the elements that comprise the stimulus and response on the initial test are also present on the final test, but with the cue and response roles of those elements reassigned. An example is Carpenter, Pashler, and Vul (2006), in which subjects first studied a set of paired associates (e.g., *beach, blanket*). They next

practiced recall of one word from each paired associate (e.g., *beach*, *?*), and on the final test were tested on the reverse case (e.g., *?*, *blanket*). Another example is Pan, Gopal, et al. (2015), in which subjects took initial tests on one term of a multi-term fact (e.g., *Overlord, an operation led by Eisenhower, began with the invasion of WHERE?*"), and on the final test had to recall a different term (e.g., "*Overlord, an operation led by WHOM, began with the invasion of Normandy?"*).

Studies in this category fall into one of four subtypes: *paired associates* (as in the aforementioned Carpenter et al. example), *triple associates* (e.g., training on a word triplet such as "*gift*, *rose*, *wine"* via "*gift, rose, ?",* and later being assessed on "*?, rose, wine"* as in Pan, Wong, et al., 2016), *multi-term facts* (as in the Pan, Gopal, et al., 2015 example), and *term-definition facts* (e.g., training on "*Vision is the ability to see"* via *"WHAT is the ability to see?"* and later being assessed using the question, *"Vision is WHAT?"*, as discussed in Pan and Rickard, 2017). For further discussions of stimulus-response rearrangement, see McDaniel, Thomas et al. (2013); Pan, Wong, et al. (2016); and Rohrer et al. (2010).

Untested materials seen during initial study. In this category, the final test assesses information that was initially studied but neither tested nor otherwise re-exposed during training. An example is Nungester and Duchastel (1982), in which subjects read a historical text passage (a 1,700-word text titled "The Victorian Era"), followed by an initial test on some aspects of that passage (e.g., "*What nationality was Prince Albert?"*). The final test then assessed other parts of the passage that were not trained (e.g., "*Where was the Crimean War?*"). Although similar to stimulus-response rearrangement in that the final test assesses a previously seen but untrained response, this category is unique in that the final test questions were not presented or tested in any form during training.

A notable characteristic about the literature on transfer to untested materials is that its

constituent studies vary greatly in the degree of semantic relatedness between tested and untested materials. In some cases, closely related principles, facts, or details are assessed (e.g., in Chan et al., 2006, an initial test question was *"The largest toucan species is?"* and the final test question was "*The most colorful toucan species is?"*; both questions referred to (relatively) highly related information that was located in adjacent portions of an initially studied article). By contrast, in other studies there is no obvious relation between initial and final test questions (outside of the fact that both stem from the same general source, as in the aforementioned Nungester and Duchastel example). However, the degree of "relatedness" between initial and final test questions defies simple categorization (we considered but dropped the use of sub-categories ranging from "same or linked concepts" to "generally related" content; for an attempt using Latent Semantic Analysis, see Chan et al., 2006). The similarity issue is further discussed in Cranney et al. (2009); Hamaker (1986); Little (2011); and Wooldridge et al. (2014).

Application and inference questions. In this category, the final test requires learners to relate prior learning to new but conceptually related information (*application*), such as a new example, scenario, or goal (Brookhart, 2015; Mayer, 2009), or to integrate prior learning in a new way but not typically with new information (*inference*), such as having to uncover (i.e., infer) a general principle (McNamara & Kintsch, 1996). In some cases, a mixture of application and inference questions is used. An example is Johnson and Mayer (2009), in which subjects took tests after watching a multimedia presentation on lightning formation (e.g., *"Please write down an explanation of how lightning works"*), followed by a final test featuring application and inference questions on that topic (e.g., *"What could you do to decrease the intensity of lightning?"*; "*Suppose you see clouds in the sky but no lightning; why not?"*). Another example is McDaniel, Howard, et al., (2009; Experiment 2), in which subjects freely recalled as much as they could remember about a text passage on brakes, followed by a final test featuring

application (e.g., *"What could be done to make brakes more effective?"*) and inference (e.g., *"Why do brakes get hot?"*) questions. By our analysis, studies in this category have featured application questions only, inference questions only, a mix of application and inference questions, or questions combining both types. That analysis relied on the definitions stated here, which we developed due to the fact that this category lacks a common definition of an application question or an inference question, and which we based on examination of question types, their descriptions in articles, and comparisons with other relevant literatures.

Different types of application questions include *analysis or evaluation* (i.e., interpreting new data or an example in the context of prior learning, such as by identifying the most appropriate concept that matches that example), *comparison or contrast* (i.e., determining similarities or differences between new data and prior learning), *prediction* (i.e., determining how a system is affected by a new situation), *redesign* (i.e., modifying a system to achieve a new goal), and *troubleshooting* (i.e., diagnosing a malfunction in a system). Different types of inference questions include *bridging inferences* (i.e., integrating multiple pieces of information that were presented separately), *conceptual inferences* (i.e., uncovering an underlying or overall principle), *elaborative inferences* (i.e., determining an implied step or component), *rhetorical inferences* (i.e., determining a main argument or thesis), and other types. For further discussion of application and/or inference question types, see Brookhart (2015); Gasparinatou and Grigoriadou (2013); Marzano, Pickering, and Pollock (2005); Mayer (2001); and McNamara and Kintsch (1996).

Problem-solving skills. In this category, after using tests to train on a multi-step problem-solving procedure, a final test involves recall and execution of that procedure to solve a new but related problem. Studies in this category fall broadly into two main sub-categories: (a) *medical diagnosis and treatment* and (b) *worked examples*. An example of the former is

Kromann, Jensen, and Ringsted (2009), in which medical students studied and took initial tests on cardiac resuscitation procedures, and then used those procedures to address similar patient scenarios with modified demographics and/or symptoms on a transfer test (e.g., "*You're about to establish I.V. access when your patient, a 75 yr old man, becomes unresponsive. You are now required to manage this patient.*"). An example of the latter is van Gog, Kester, and Paas (2011), in which students studied worked examples (i.e., a problem in which the solution steps are shown, providing a step-by-step guide on how to arrive at the correct solution) of circuit troubleshooting problems, took practice tests on those problems (e.g., *"Determine how this circuit should function using Ohm's law"*), and then attempted to solve new problems with different values and often greater complexity (e.g., again determining how a new circuit should function, but in this instance there are two circuit faults rather than one as seen previously). While similar to application questions in that new information is commonly presented on the transfer test (and, in some cases, malfunctions in a system need to be identified), the problem-solving category is unique in that training is focused on learning a sequence or set of to-be-executed procedures. The types of problems used in this category are further discussed in Karpicke and Aue (2015); Larsen, Butler, and Roediger (2008); Leahy, Hanham, and Sweller (2015); Rawson (2015); and van Gog and Sweller (2015).

Mediator and related word cues. In this category, after training on paired associate words or word lists via a cued recall test, a final cued recall test involves recall of the same words but in response to different (i.e., mediator or related word) cues. An example is Carpenter (2011; Experiment 2), in which subjects took cued recall tests without feedback on paired associates (e.g., *mother, ?* for which the answer is *child*), followed by a final test in which they again attempted to retrieve target words, but in response to mediator word (e.g., *father, ?*) or related word (e.g., *birth, ?*) cues. *Mediator cues* are words that have strong preexisting semantic

associations with cues (e.g., *father* is a mediator for the cue *mother*), whereas *related cues* are words that are weakly related to targets (e.g., *birth* is related to the target *child*). For further discussion of this transfer category, which is the newest and least populated of those analyzed in this review, see Cho, Neely, Brennan, Vitrano, and Crocco (2016); Coppens et al. (2016); and Rawson, Vaughn, and Carpenter (2015).

Other transfer contexts. Besides the six major categories, the literature also contains studies exploring other types of transfer, but in all such cases there are currently too few papers to include in the present meta-analyses (however, in several of these studies, one or more of the aforementioned transfer categories was also explored; those results are included in the meta-analyses). There are two categories with two or more articles: transfer of *category learning* (wherein subjects learn to classify visually presented category exemplars; e.g., Baghdady, Carnahan, Lam, & Woods, 2014; Jacoby, Wahleim, & Coane, 2010), and transfer of *visuospatial learning* (wherein subjects recall locations and/or make route or directional judgments; e.g., Carpenter & Kelly, 2012; Rohrer et al., 2010, Experiment 2). Other studies are the first and, as of this writing, only investigation of yet other types of transfer (e.g., Kang, McDaniel, & Pashler, 2011, which involved training on mathematical functions; see also George & Wiley, 2016, involving analogical transfer). Finally, there are additional transfer contexts (e.g., changes in social contexts, as in individuals vs. groups) that have yet to be investigated in this literature to date.

Candidate Effect Size Moderators

Various candidate moderators of testing and/or transfer effects have been catalogued in the test-enhanced learning and broader transfer literatures. In the test-enhanced learning literature, these fall into three categories: (a) *encoding factors* (e.g., the number of training trials per item, the presence or absence of feedback, the type of feedback provided, the initial test

format, and proportion correct on the initial test), (b) *retrieval factors* (e.g., the types of final test questions), and (c) *other design variables* (e.g., the length of the retention interval between the initial and final tests, as well as the type of subject materials that are being learned). Any of these factors may influence test-enhanced learning, and by extension may also influence testing's ability to yield positive transfer. With regard to (a), the use of increased training trials, correct answer feedback (and especially more extensive feedback, such as feedback containing explanations), more difficult initial test formats (e.g., cued recall rather than recognition), and relatively high initial test performance (e.g., > 0.50 proportion correct) have been associated with larger testing effects; many of these factors have also been hypothesized to improve transfer (for data and discussions see Butler, Godbole, & Marsh, 2013; Dunlosky et al., 2013; Goode, Geraci, & Roediger, 2008; Jensen, McDaniel, Woodard, & Kummer, 2015; Karpicke & Aue, 2015; McDaniel & Masson, 1985; McDaniel, Thomas, et al., 2013; McDaniel, Wildman, et al., 2012; and Rowland, 2014). With regard to (b), the use of more difficult final test formats (e.g., cued recall rather than recognition) can yield larger testing effects (Halamish & Bjork, 2011; Rowland, 2014); corresponding effects on transfer (as well as those of other retrieval factors) have yet to be thoroughly investigated. Regarding (c), the testing effect tends to be larger at retention intervals of one day or more relative to shorter intervals (Rowland, 2014), a pattern that may also hold for transfer, whereas the role of subject materials on testing or transfer effects has been the subject of differing hypotheses (e.g., Karpicke & Aue, 2015; Pan, Gopal, et al., 2015; van Gog & Sweller, 2015).

Correspondingly, in a widely-cited review of the broader transfer literature, Gick and Holyoak (1987; see also Barnett & Ceci, 2002; Brooks & Dansereau; Haskell, 2001; McGeoch, 1942; Perkins & Solomon, 1994; Singley & Anderson, 1989) concluded that four types of factors moderate transfer: (a) the *structure* of the training and transfer tasks (e.g., the type of knowledge

that needs to be learned and how similar the tasks are to one another), (b) *encoding factors* (e.g., the number and variability of examples provided during training, amount of training, types of instructions given during training, and degree of abstract learning during training), (c) *retrieval factors* (e.g., whether learners are informed of the transfer context, the similarity of the transfer cues to those seen during training, and the similarity of responses on the transfer task to those made during training, and (d) *prior knowledge* and other pre-experimental factors. With regard to (a), the more structurally similar the training and transfer tasks are, the more likely transfer is expected (Haskell, 2001), although transfer may also generally vary by knowledge type (Healy, 2007). With regard to (b), an increased number and variety of examples, more training, instructions to learn underlying principles, and more abstract learning have all been associated with improved transfer (Gick & Holyoak, 1980, 1987; Haskell, 2001; Perkins & Solomon, 1994). Regarding (c), increased similarity in cues and/or responses between the training and transfer tasks (Wylie, 1919; Osgood, 1949), as well as the provision of hints (Gick & Holyoak, 1980, 1987), have been associated with improved transfer. Regarding (d), if relevant to the transfer context, prior expertise may also increase transfer (Gick & Holyoak, 1987; Haskell, 2001).

Candidate moderators investigated in the meta-analyses. Drawing from both literatures and our observation of potentially important design factors during the literature review process, each effect size included in this review was coded with respect to seven potentially applicable and analyzable candidate moderators. These candidate moderators were: (a) *between- vs. within-subjects design,* (b) *number of training phase item repetitions*, (c) *initial test performance* (i.e., proportion correct during the training phase)*,* (d) *retention interval*, (e) *correct answer feedback*, (f) *elaborated retrieval practice*, and (g) *response congruency.* In some cases, other previously hypothesized moderators could not be analyzed due to their being too rarely or not at all investigated in this literature. Each candidate moderator was coded by the authors; for

purposes of intercoder agreement and verifying accuracy, a subset of papers was also separately coded a second time by a trained research assistant. For candidate moderators (a) to (e), all published articles were coded a second time; for candidate moderators (f) and (g), a randomly selected 25% of the overall dataset (corresponding to 48 effect sizes; cf. Bujang & Baharum, 2017) was coded again. Any discrepancies between raters were resolved by discussion and arriving at mutual agreement.

Each of the seven candidate moderators was investigated in the overall meta-analyses. Where there were sufficient data to do so, they were also investigated in the category-level meta-analyses. The candidate moderators were defined as follows.

Between- vs. within-subjects design. With respect to training condition (e.g., testing vs. non-testing reexposure control), each study was coded as using a between- or within-subjects design. In the test-enhanced learning literature, between-subjects designs typically yield larger effect sizes (Rowland, 2014), although that result has not always been obtained when between- vs. within-subjects group assignment has been manipulated within a single experiment (e.g., Huff, Balota, & Hutchison, 2014; Rowland, Littrell-Baez, Sensenig, & DeLosh, 2014; Experiment 3; Soderstrom & Bjork, 2014). Intercoder agreement was $\kappa = 0.97$.

Number of training phase item repetitions. Each study was coded for the number of repetitions of each item during training, ranging from one to five, as a continuous variable. In all meta-analyzed studies, the number of item repetitions in the testing and non-testing reexposure control conditions were identical. Some researchers have emphasized that repeated testing is an important factor in maximizing the benefits of test-enhanced learning and yielding transfer (e.g., McDaniel, Thomas, et al., 2013; McDaniel, Wildman, et al., 2012). Intercoder agreement (proportion of identically extracted values) was 1.00.

Initial test performance. Where reported, initial test (i.e., training test) proportion

correct data was recorded as a continuous variable. Initial test performance, particularly in the absence of correct answer feedback, has been suggested as a moderator of test-enhanced learning (e.g., Kang et al., 2007; Rowland, 2014; Smith & Karpicke, 2014). Where multiple initial test repetitions were administered, proportion correct data from the last of those tests was included in the analyses. Initial test performance was the only candidate moderator for which there was missing data (i.e., for 30% of included effect sizes, that data was not collected or reported). Intercoder agreement (proportion of identically extracted values) was 1.00.

Retention interval. The length of time between the end of training and the final test, in hrs,, was recorded for each study as a continuous variable. In the test-enhanced learning literature, the magnitude of the testing effect tends to become larger as the duration of the retention interval increases (e.g., Carpenter, Pashler, et al., 2008; Roediger & Karpicke, 2006), with retention intervals longer than one day often yielding larger testing effects than retention intervals that are shorter than one day (Rowland, 2014). Intercoder agreement (proportion of identically extracted values) was 0.98.

Correct answer feedback. Each study was coded for the presence or absence of correct answer feedback during training. All experiments in which subjects were able to view the correct answers to initial test questions shortly after answering them (i.e., in the same training session) were coded (as the value 1) as providing correct answer feedback (Rowland, 2014). Cases with no feedback were coded with the value of zero. Feedback that did not include exposure to all the correct answers was coded as no feedback, of which there were two types: (a) feedback involving the number of questions scored correctly out of the total number of questions (as in Meyer & Logan, 2013), and (b) feedback provided during an instructor-led brief discussion session that was general in nature and did not specifically address individual subjects' responses (as described in Kromann et al., 2009; Kromann, Jensen, & Ringsted, 2010; Kromann,

Bohnstedt, Jensen, Ringsted, 2010). One study in which data from feedback and no feedback conditions were collapsed together (Butler & Roediger, 2007) was excluded from analyses involving correct answer feedback. The emphasis on the correct answers being presented during feedback is due to the fact that feedback lacking such information is often no better than no feedback at all (e.g., Anderson, Kulhavy & Anders, 1971; Kulhavy & Anderson, 1972; Pashler, Cepeda, Wixted, & Rohrer, 2005; for an exception, see Butler, Karpicke, & Roediger, 2008). Intercoder agreement was $\kappa = 0.97$.

Elaborated retrieval practice. Each study was coded for the presence or absence of several retrieval-specific and post-retrieval training manipulations – which we classified as *broad encoding methods* and *elaborative feedback*, respectively – that have been previously been hypothesized in the literature to enhance transfer. Studies in which either type of training manipulation was utilized, or both (e.g., Little, 2011, Experiment 5, and McDaniel et al., 2015), were classified as using elaborated retrieval practice. We originally planned on analyzing both broad encoding methods and elaborative feedback separately, but report analyses of the two in combination – as the *elaborated retrieval practice* candidate moderator (which is not to be confused with elaborative retrieval hypothesis) – due to insufficient data at most category levels when those components were fitted separately. If broad encoding, elaborative feedback, or both (as defined below) was present, this variable was coded as 1; otherwise it was coded as zero. Intercoder agreement was $\kappa = 0.92$.

Broad encoding methods. Each study was coded for the presence or absence of initial cued recall or multiple-choice tests that directed subjects to specifically think about additional information (e.g., content related to the tested concept or target item) or to retrieve multiple pieces of information that pertain to a given concept or target item while making a response or responses. This contrasts with far more common cued recall and multiple-choice initial tests

which involve retrieval of (or recognition of) a single response for a given concept or target item and/or do not directly specify the consideration of additional information while making responses. The majority of included studies did not feature broad encoding methods, but several (specific conditions identified in Table 1) involved one or more of the following four techniques:

1. *Broad retrieval instructions* – recalling any and all related information that was presented during the initial study phase before responding (e.g., "think of everything you can recall that is possibly related to the answer"; as occurred in Chan, et al., 2006; Experiment 3).
2. *Discrimination instructions* – deliberating on each of a set of provided multiple-choice answer options prior to selecting an answer (as occurred in Little, 2011; Experiment 5; those answers would later be referenced on a transfer test).
3. *Explanatory recall instructions* – constructing a detailed explanation, in one's own words, while responding on an initial test (as occurred in Hinze et al., 2013, Experiment 3; subjects were provided topic sentence prompts).
4. *Use of high and low order questions* – answering multiple questions for a given concept or fact, with those questions not just involving pure recall, but also involving higher order cognitive processes (to use the terminology of Bloom's taxonomy, namely questions requiring the learner to evaluate, analyze, or synthesize; cf. Anderson & Krathwohl, 2001; Bloom, 1956, 1984). An example is McDaniel, Bugg, et al. (2015), in which each concept was trained with application questions and term retrieval questions.

Prior discussions of the potential effectiveness of these methods at yielding positive transfer can be found in Chan et al. (2006); Hinze et al. (2013); Jensen et al. (2014); Little (2011); Nguyen and McDaniel (2016); and Pan and Rickard (2017).

We acknowledge that the definition of broad encoding methods presented here may not receive total agreement from readers. In particular, free recall tests might be argued as a method

that also induces broad encoding of to-be-learned materials. However, due to the ambiguity that results from having completely open-ended answers, the production of which may or may not involve processing of information that is later relevant to a transfer test, studies using free recall were not coded as involving broad encoding methods. Accordingly, it should be emphasized that any conclusions that can be drawn from the current meta-analyses regarding broad encoding methods apply only to the techniques described in this section.

Elaborative feedback. Each study was coded for the presence or absence of post-retrieval activities that, beyond the processing of brief correct answer feedback, enabled subjects to extensively restudy target materials or information that would later be relevant to correct responding on a transfer test. Studies featuring elaborative feedback (specific experimental conditions are identified in Table 1) used one (or more) of the following three methods:

1. *Post-retrieval restudy opportunities beyond simple correct answer feedback*. This includes re-reading of entire text passages (e.g., McDaniel, Howard, et al., 2009; Wooldridge et al., 2014; Zhou et al., 2013) or extensive review in preparation for an upcoming high-stakes test (e.g., in a classroom setting, as occurred over a week-long period in Balch, 1998). These methods have been, by far, the most common implementations of elaborative feedback in the literature.
2. *Explanatory feedback*. Feedback that, beyond simply providing a brief (e.g., one word or short phrase) correct answer, contains an explanation of the answer, an explanation of the underlying concept, and/or reasons why it is correct; that feedback (usually comprised of several sentences) would later be relevant on a transfer test (e.g., the conceptual questions conditions of Butler, 2010; see also McDaniel, Wildman, et al., 2012).
3. *Extended and detailed feedback* – feedback that could be repeatedly viewed after testing for an unlimited period of time, including test questions, responses, and correct answers

(e.g., McDaniel, Anderson, et al., 2007; students could review that feedback for a week), as well as feedback that includes all target materials (e.g., complete sets of premises used to form logical scenarios as in Eglington & Kang, 2016).

Prior discussions of the potential effectiveness of these methods at yielding positive transfer can be found in Butler et al. (2013); McDaniel, Howard, et al. (2009); McDaniel and Little (in press); Pan, Gopal, et al. (2015); Pan and Rickard (2017); and van Eersel et al. (2016).

Coding of studies for elaborative feedback was performed through inspection of article methods sections and, in some cases, experimental materials. We acknowledge that the definition of elaborative feedback presented here may not fully align with some readers' views (for related discussions of feedback types, see Butler et al., 2013; Kulhavy & Stock, 1989). Ultimately, as with broad encoding methods, the classification of studies for elaborative feedback reflected the available evidence and (in cases of ambiguity) our judgment (with both authors arriving at mutual agreement). Any conclusions about elaborative feedback apply only to the techniques described in this section.

It should also be noted that elaborative feedback was provided in 17 of 23 included effect sizes involving initial free recall tests (in the form of post-retrieval restudy opportunities, a common implementation of which involved a free recall attempt, a restudy period, and then a second free recall attempt). Thus, the coding of studies for elaborative feedback (and hence elaborated retrieval practice) encompassed the majority of studies in the literature featuring free recall on the initial test, plus addressed the common training method of free recall testing accompanied by restudy opportunities.

Response congruency. Each study was dichotomously coded for the presence or absence of correct answer response congruency on the initial and final test. Response congruency was defined as having the same or very substantially overlapping answers. We classified studies

based on descriptions or examples of the materials and methods used for initial and final tests (as available in the source articles). When performing this classification, we used a stringent criterion wherein ambiguous cases were treated as not congruent; only unambiguous and obvious cases where the same or substantially overlapping answers were present on initial and final tests were classified as having response congruency. Because of this stringent criterion, any significant effects for response congruency that may be observed in the current meta-analyses may actually understate its influence.[3] Intercoder agreement was $\kappa = 0.82$.

Response congruency generally held in some categories and sub-categories. This included (a) the majority of studies involving transfer across test formats (i.e., all studies in the cued recall to free recall, cued recall to multiple-choice, and multiple-choice to cued recall sub-categories), (b) all studies involving transfer to mediator and related word cues, and (c) a sub-category of transfer of problem-solving skills, namely medical diagnosis and treatment (wherein initial and final tests involve scenarios and procedures that are "essentially the same"; per Kromann et al., 2009, p. 23). Together, these studies accounted for 40% of all effect sizes in the meta-analyses. Conversely, for studies of transfer to stimulus-response rearrangement, all correct responses on the final test were, by definition, different from the correct responses on the initial test. Similarly, there was minimal-to-no response congruency for studies of transfer to untested materials seen during initial study, in nearly all studies involving transfer to application and inference questions, and in all studies in the worked examples sub-category of transfer of problem-solving skills. Exceptions in the application and inference category were McDaniel,

[3] Having strong response congruency might seem to be antithetical to the definition of transfer (i.e., the same response does not constitute a new context). However, as an example, one could be asked a definitional question (e.g., *"The degree to which a measurement or a test is consistent is called…?",* for which the answer is *reliability*), and later be asked to provide the same response for an application question (e.g., *Jon weighed a stone on the same scale three times and obtained different readings each time; the scale lacks…?*), which is a case of the same response used in a different context (i.e., a clear case of transfer). There are a variety of such cases.

Bugg et al. (2016) and Nguyen and McDaniel (2016; Experiment 1), in which the correct responses to final transfer test questions were identical to correct responses during training.

For transfer across test formats, the sub-categories involving free recall on the initial test were regarded as having ambiguous response congruency (due to the fact that information that is retrieved on a free recall test may or may not match the answers on a subsequent test that involves more precisely specified cues and responses), and hence were coded as having no congruency. Similarly, all studies in the cued recall to recognition sub-category were rated as having no response congruency, with the exception of Carpenter (2011, Experiment 1). In that sub-category, the recognition test typically involves making old/new judgments to previously seen (i.e., old items) and new (i.e., lure items) stimuli; the previously seen stimuli have strong response congruency, whereas new stimuli do not (final test data was separately reported for old items only in the case of the Carpenter study; those particular results were coded as having response congruency).

Additional candidate effect size moderators. Four additional category-specific candidate moderators, all involving comparison of sub-categories within a single category, were also analyzed due to prior treatment or speculation in the literature. These were: *free recall vs. not free recall on the initial test* and *multiple-choice vs. not multiple-choice on the initial test* (test format category); *paired associates vs. non-paired associates* (stimulus-response rearrangement category); and *worked examples vs. medical diagnosis and treatment* (problem-solving skills category).

Computing Effect Sizes, Sampling Variability, and Confidence Intervals

All formal meta-analyses were conducted using a standardized effect size, Cohen's *d*. Each effect size was computed as the mean difference over subjects in proportion correct between the two final test conditions of interest (transfer condition minus non-testing reexposure

control condition) divided by the standard deviation of that mean difference. All effect sizes were calculated from a reported or derivable *t* statistic or a single degree of freedom *F* statistic, and a reported or derivable sample size or degrees of freedom. For between-subjects designs, *t* values were converted into *d* using the equation (Glass, McGaw, & Smith, 1981):

$$d = t\sqrt{\left(\frac{n_T + n_R}{n_T n_R}\right)}, \tag{1}$$

where n_T refers to the sample size in the test condition and n_R refers to the sample size in the non-testing reexposure control condition. For between-subjects designs in which the results of a statistical test of interest was not reported, but in which group-level means, their associated standard errors or deviations, and sample sizes were available or derivable from graphically reported information (see Pan & Rickard, 2015, for a description of the method of extracting data values from graphically presented information), the *t* statistic was calculated from those data prior to effect size conversion.

For within-subjects designs, *t* values were converted to *d* using the equation (from Dunlap, Cortina, Vaslow, & Burke, 1996):

$$d = t\sqrt{\frac{2(1 - r)}{n}}, \tag{2}$$

where *n* is the sample size and *r* is the estimated correlation of final test performance in the testing and non-testing reexposure control conditions. Following Dunlap et al. (1996) and Rowland (2014), the unknown value of *r* was set to .5 with the expectation of making within- and between-subjects effect sizes roughly comparable, thus reducing Equation 2 to:

$$d = \frac{t}{\sqrt{n}}. \tag{3}$$

All effect sizes were computed using Equations 1 or 3 regardless of whether an effect

size was directly reported in the article text. We elected to do so for consistency and accuracy, given that (a) less than one quarter of included studies reported any effect size information, and (b) not all of the reported effect sizes matched those that were calculated using the above methods (our calculations produced effect sizes that matched within $d = \pm 0.05$ for 30 of the 50 reported effect sizes). Given the long history in psychology of statistical training on *t* and *F* statistics, effect size estimates based on those statistics may be more accurate than those reported (for related discussion, see Lakens, Hilgard, & Staaks, 2016).

The effect size sampling variability (*sv*) was calculated from equations specified in Morris and DeShon (2002). For within-subjects designs, *sv* was calculated using the equation:

$$sv = \left(\frac{1}{n}\right)\left(\frac{n-1}{n-3}\right)\left(1 + nd^2\right) - \left(\frac{d^2}{c^2}\right), \quad (4)$$

where *c* refers to the bias function in Hedges (1982) that is calculated using the equation:

$$c = 1 - \left(\frac{3}{4df - 1}\right), \quad (5)$$

with *df* referring to degrees of freedom. For between-subjects designs, *sv* was calculated using the equation (from Morris & DeShon, 2002):

$$sv = \left(\frac{1}{\tilde{n}}\right)\left(\frac{N-2}{N-4}\right)\left(1 + \tilde{n}d^2\right) - \left(\frac{d^2}{c^2}\right), \quad (6)$$

where *N* is the combined sample size of both groups, and $\tilde{n} = (n_T n_R)/(n_T + n_R)$.

Ninety-five percent confidence intervals for each effect size in the forest plots of Figure 2 were computed using IBM SPSS Statistics (International Business Machines Corp., Armonk, NY) software and Smithson's (2003) publicly available noncentral *t* scripts. This method uses the software's noncentral *t* calculator and Laubscher's (1960) normal approximation method (Wuensch, 2012).

Random-Effects Meta-Analysis with Robust Variance Estimation

Random-effects meta-analysis (Borenstein, Hedges, Higgins, & Rothstein, 2010; Raudenbush, 2009) was one of two main approaches employed for quantitative meta-analyses. Two random effects, study and experiment (within study), were estimated hierarchically using the following model:

$$T_{ij} = X_{ij}\beta + \theta_i + \eta_{ij} + \varepsilon_{ij}, \quad (7)$$

where T_{ij} is the estimated effect size for group i in study j, X_{ij} is the design matrix in study j, β is the vector of regression coefficients, θ_i is the study-level random effect, η_{ij} is the group-level random effect, and ε_{ij} is the sampling error.

In random-effects meta-analysis the observed effect size at each level of the hierarchy (i.e., for each study and each group within study) is treated as a random deviate from its own population effect size distribution. The degree to which the effect sizes are in fact heterogeneous (i.e., random deviates from different distributions) vs. homogeneous (i.e., random deviates from the same distribution) can be quantified, both prior to and after fitting candidate moderator variables. In the current model, the residual variation of the effect size estimate T_{ij} can be decomposed as:

$$V(T_{ij}) = \tau^2 + \omega^2 + \upsilon_{ij}, \quad (8)$$

where τ^2 is the variance of the between-study residuals, θ_i, and ω^2 is the variance of the within-study residuals, η_{ij}, and υ_{ij} is the known sampling variability of each group. Estimates of τ^2 and (or) ω^2 that are greater than zero suggest that heterogeneity is present and that fixed-effects moderator variables (i.e., meta-regression) may help explain differences in effect sizes over papers and (or) groups within papers.

Given that the covariance structure of the effect size estimates is unknown in the transfer

of test-enhanced learning literature, we employed robust variance estimation (Hedges, Tipton, & Johnson, 2010; Tanner-Smith & Tipton, 2013) in the model fitting. All meta-analyses were performed using Stata (StataCorp LP, College Station, TX, USA) and the macro *robumeta.ado*, which can be downloaded from the Stata Statistical Software Components archive (SSC).

Possible effect size dependencies in our analyses mainly consisted of multiple experiments within a study (i.e., within a single paper). Accordingly, we used the macro's hierarchical weight type option, which accommodates such dependencies (see Tanner-Smith & Tipton, 2013).

Analyses Adjusted for Publication Bias

Although the hierarchical random-effects meta-analysis described above provides relatively good statistical power and can accommodate dependencies due to nesting of experiments within paper, it does not adjust for potential publication bias. As supplemental analyses that can both detect and adjust for publication bias, we used the precision effect estimate with standard errors (PET-PEESE) method (Stanley & Doucouliagos, 2014) and the sensitivity analysis with a priori weight functions (Vevea & Woods, 2005) approach.

PET-PEESE analyses. Both the PET and PEESE analysis procedures involve standard, weighted linear regression in which *d* is predicted by either the standard error (PET) or the *sv* (PEESE), weighted by study precision, 1/*sv*. Any significant moderator variables that were identified in the random-effects analysis were also included in the PET-PEESE analyses. The general equation for PET (Stanley & Doucouliagos, 2014) is,

$$d = \beta_0 + \beta_1 se_i + \Sigma_k \alpha_k z_k + \varepsilon_i, \quad (9)$$

and for PEESE is,

$$d = \beta_0 + \beta_1 sv_i + \Sigma_k \alpha_k z_k + \varepsilon_i, \quad (10)$$

where β_0 refers to the intercept estimate, β_1 is the slope estimate, se_i is the standard error for the ith effect size, sv is the corresponding sampling variability, z_k is the k^{th} moderator, and α_k is the corresponding effect size estimate.

If an effect is suspected to be entirely due to publication bias, then Equation 9 is appropriate. If an effect is believed to be genuine, then Equation 10 is appropriate. Stanley (2017) recommends the initial application of Equation 9. If the intercept is positive or is non-significantly negative at $p > .10$, then performing the primary analysis using Equation 10 is recommended. We used that criterion, along with the additional criterion that PEESE was used if one or more of the moderators that were identified in the random-effects analysis was significant at the 0.05 level (all dichotomous moderators were coded as in the random-effects analysis such that the level of the moderator that was expected to yield smaller effect sizes took a value of zero and the level expected to yield larger effect sizes took a value of 1). By those criteria, PEESE rather than PET was in all cases indicated, and thus only the PEESE results are described in the Results section.

If there is publication bias, then it is expected that there will be a positive slope (β_1) relating d to sv. That effect is expected because studies with low precision will have the highest variability in effect size estimates, and because in most cases unpublished studies are those with low precision and small effect size estimates that do not reach the traditional statistical significance level. Effect size estimates adjusted for publication bias are assessed at the intercept (including moderator intercepts), corresponding to the hypothetical best case study with zero sampling variability.

In our view, the hierarchical random-effects and PEESE analyses are complementary. Accordingly, confidence is highest when an effect that is detected in the former analysis method is also apparent in the latter.

Sensitivity analyses. We performed sensitivity analysis using the weight-function approach developed by Vevea and Woods (2005; see also Vevea & Hedges, 1995) to estimate the consequences of different degrees of possible publication bias on meta-analytic outcomes. Vevea and Woods specify a set of fixed weights that can be used to explore four different scenarios of possible publication bias: the cases of moderate and severe bias for both one- and two-tailed tests. We conducted sensitivity analyses using Vevea and Woods' publicly available sensitivity analysis tool (available at: https://vevealab.shinyapps.io/WeightFunctionModel/) for each of those four scenarios and using the authors' example *p*-value cutoffs of .001, .01, .05, and .50 (and for the two-tailed cases, also .999, .99, and .95).

Results

The meta-analyses are reported in the following order. First, we report random-effect meta-analyses on the entire dataset of 192 effect sizes across 67 papers. This includes (a) the estimated weighted mean effect size for the transfer of test-enhanced learning literature, (b) fits of each candidate moderator in isolation, and (c) simultaneous fits of candidate moderators (i.e., meta-regression analysis), along with selection of a subset of statistically significant moderators, yielding a final model. The latter analysis involved fitting all candidate moderators and then iteratively eliminating the least significant (i.e., largest *p*-value) candidate moderator from the model, one at a time (Van den Bussche, Noortgate, & Reynvoet, 2009), until all remaining moderators were statistically significant or marginally significant (here and throughout this review, we used a significance criterion of $\alpha = 0.05$). Marginally significant moderator variables were included because it may be of interest to the field to know about factors that may influence transfer, even though the evidence is currently weak.

Second, we report the results of PEESE and sensitivity analyses for the entire dataset. In these analyses, all moderator variables that were identified in the random-effects meta-analyses

were included. Third, we repeat the same analysis sequence, but limited to the 135 effect sizes for which proportion correct on the initial test was reported (and hence that candidate moderator could be tested). Initial test performance was the only candidate moderator with substantial missing data in our dataset. Fourth, we report category-level meta-analyses where possible, each of which involved the same general analysis sequence as in the overall meta-analyses (i.e., random-effects meta-analyses involving steps a, b, and c, after which the results of PEESE and sensitivity analyses are reported).

Overall Meta-Analyses

Across the entire dataset of included studies, the weighted mean effect size in the random-effects model was $d = 0.40$ ($p < .00001$), a medium effect by traditional standards (Cohen, 1988), with a 95% confidence interval (C.I.) of [0.31, 0.50]. That result supports, for the first time at the literature level through quantitative means, the general view that test-enhanced learning can yield transfer (pending tests for publication bias). There was also evidence of heterogeneity, however, both at the between paper ($\tau^2 = 0.084$) and the experiments within paper ($\omega^2 = 0.050$) levels. It is perhaps not surprising that heterogeneity is present in this analysis, given the wide range of materials and transfer contexts in the literature.

Further insight into the effect size patterns can be gained by visual inspection of the forest plots in Figure 2 (all panels). Those plots display effect sizes and confidence intervals for each experiment organized by the six major transfer categories and their sub-categories. In the context of the overall pattern of positive transfer, there is substantial variability in effect sizes between both major categories and sub-categories (reflected quantitatively in the τ^2 value) and, in some categories, over effect sizes within paper (reflected in the ω^2 value). The weighted mean effect sizes and confidence intervals for the entire dataset and for each category is shown in Table 2.

(Table 2 around here)

Single and simultaneous moderator fits to the full dataset using random-effects meta-analysis. The heterogeneity observed in the overall model fit motivates consideration of candidate moderators. These are described next.

Single moderator fits. Results of the single moderator fits to the overall dataset are listed in Table 3. In those fits, only response congruency was significant ($p = .0006$). When there was no response congruency between the initial and final tests, the weighted effect size was $d = 0.28$, $p < .00001$, C.I. [0.17, 0.39]; if response congruency held, the estimated effect size increased by $d = 0.30$, $p = .0006$, C.I. [0.14, 0.47], yielding an estimated effect size of $d = 0.58$.

(Table 3 around here)

Simultaneous moderator fits. Results of the simultaneous moderator fits to the overall dataset are listed in Table 3. In the final model, the moderators of response congruency and elaborated retrieval practice were significant ($ps \leq .0094$). No other candidate moderators approached significance in that model. Having the same correct responses on the initial and final tests yielded a higher estimate of positive transfer (estimated increase of $d = 0.35$, $p = .0002$, C.I. [0.18, 0.51]), as did the use of elaborated retrieval practice (estimated increase of $d = 0.22$, $p = .0094$, C.I. [0.059, 0.38]). When both of those factors were present, the estimated transfer effect size was $d = 0.78$. When neither was present, the estimated transfer effect size was $d = 0.21$. Between-paper heterogeneity was reduced more than within-paper heterogeneity in that model ($\tau^2 = 0.058$, $\omega^2 = 0.042$).

Analyses adjusted for publication bias. Results of PEESE analyses to the full dataset are listed in Table 4 (for completeness, the table includes the cases in which the previously identified moderators are or are not included, although we focus on the results for the former case). There was a highly significant effect ($p < .0001$) of *sv,* suggesting publication bias. Moreover, unlike

the random-effects analyses, the intercept, representing the estimated effect size when neither moderator effect is present, is effectively zero. However, the effect size estimates for the moderator variables are highly consistent with the random-effects analyses discussed earlier. When response congruency is present, the estimated increase in *d* is 0.36, and when elaborated retrieval practice is present, the estimated increase in *d* is 0.18. Thus, there is no evidence that the moderator results in the final random-effects model were meaningfully contaminated by publication bias.

(Table 4 around here)

Results of sensitivity analyses using various selection methods are listed in Table 5. Under all four scenarios of potential publication bias (moderate and severe one- and two-tailed publication bias, respectively), the effect size estimates for the two moderators were similar to or larger than those observed in the random-effects and PEESE analyses. The intercept was more variable under the four scenarios and became negative in the case of severe one-tailed publication bias. It is also notable that the intercept in the PEESE analysis to the overall dataset with moderators included ($d = 0.015$) falls in between the estimated intercepts in the moderate one-tailed ($d = 0.12$) and severe one-tailed ($d = -0.12$) sensitivity analyses. In our estimation, one-tailed publication bias (i.e., publication bias that obscures cases of non-significant and negative transfer) is the more likely in this literature.

(Table 5 around here)

Meta-Analyses Involving All Studies with Initial Test Performance Data

In the random-effects analysis, the weighted mean effect size for studies in which proportion correct on the initial test was reported ($k = 135$ effect sizes; 30% of effect sizes lacking such data were excluded) was $d = 0.41$, $p < .00001$, C.I. [0.30, 0.52], which is nearly identical to that of the full dataset. Between-paper and within-paper heterogeneity was also

highly comparable (τ^2= .077, ω^2 = 0.050).

Single and simultaneous moderator fits to studies with initial test performance data using random-effects meta-analysis. Results of single and simultaneous moderator fits to the 135 effect sizes with initial test data are listed in Table 3.

Single moderator fits. In the single moderator fits (Table 3), the number of training phase repetitions, response congruency, and initial test performance moderators were significant (ps ≤ .0048). The results for response congruency matched those in the overall meta-analyses. The results for the number of training phase repetitions moderator suggests that with each added repetition during training, the proportion correct transfer effect size increases by an estimated d = 0.13, p = .0011, C.I. [0.060, 0.20]. However, as described below, that moderator did not survive in the simultaneous moderator fits.

Simultaneous moderator fits. In the final model (Table 3), there were three robust moderators: response congruency and elaborated retrieval practice (as in the overall fits to the full dataset), plus initial test performance (ps ≤ .015). Between-paper heterogeneity was substantially reduced (from τ^2 = .077 in the model fit with no moderators to τ^2 = 0.028), although within-paper heterogeneity was not (ω^2 = 0.061). Having achieved a higher proportion correct on the initial test was associated with a greater likelihood of transfer – as was the presence of response congruency and the use of elaborated retrieval practice. With regard to initial test performance, the proportion correct transfer effect size increased (Δd) by an estimated 0.0058 for every increment of 0.01 in initial test proportion correct. Across the full observed range of observed initial test proportion correct in the sample (0.19 to 0.98), the total Δd was 0.46.

Analyses adjusted for publication bias. Results of PEESE analyses for effect sizes with initial test performance data are listed in Table 4. In the analyses with moderators fitted, *sv* was again a highly significant moderator (p = .0002), suggesting substantial publication bias. The

intercept in this case was negative, suggesting negative transfer in the absence of response congruency and elaborated retrieval practice, and when proportion correct on the initial test is zero. The moderator effect sizes were again analogous to those of the random-effects analyses; the parameter estimates for response congruency and initial test accuracy (Δd = 0.25 and 0.50, respectively) were only marginally smaller than in the random-effects analysis (compare to Table 3) and they remained highly significant. The estimate for elaborated retrieval practice was smaller in this analysis but also remained positive.

Results of the sensitivity analyses are listed in Table 5. Again, the moderator effect sizes were not reduced under different scenarios of publication bias. As with the sensitivity analyses to the overall dataset, the estimated intercept was more variable under the different scenarios, especially for the case of one-tailed publication bias; the intercept for the case of moderate bias (d = ─0.26) most closely resembles the intercept in the PEESE analyses (d = ─0.30) of the initial test performance dataset. This result reinforces our earlier inference of at least moderate publication bias in this literature.

Category-Level Meta-Analyses

Results of category-level single and simultaneous moderator fits are described as follows.

Test format. A category-level random-effects meta-analysis on 56 effect sizes from all 29 included studies in the test format category (comprising six transfer sub-categories involving different combinations of initial and final test formats) yielded a weighted mean effect size of d = 0.58 (a medium-large effect), $p < .00001$, C.I. [0.43, 0.73]. This result confirms the conclusion in the literature of positive transfer across test formats (that is also evident in the forest plot in Figure 2, panel a). Heterogeneity remained, however, at both the between paper ($\tau^2 = 0.056$) and, more prominently, within-paper ($\omega^2 = 0.089$) levels.

Single and simultaneous moderator fits. We first tested the following five candidate

moderators using random-effects meta-analysis: between- vs. within-subjects design, correct answer feedback, response congruency, number of training phase item repetitions, and retention interval. Additionally, prompted by hypotheses in the literature about the relative effectiveness of different initial test formats (e.g., McDaniel, McDermott, et al., 2007), we examined two dichotomous candidate moderators involving initial test format: free recall vs. not free recall on the initial test and multiple-choice vs. not multiple-choice on the initial test. There were too few studies involving elaborated retrieval practice for that candidate moderator to be analyzed. In the single moderator fits (Table 6), only the between- vs. within-subjects moderator was significant. For within-subjects designs, the weighted effect size was $d = 0.49$, $p < .00001$, C.I. [0.33, 0.66]; the use of a between-subjects design increased the estimated effect size by $d = 0.35$, $p = .0062$, C.I. [0.12, 0.58], yielding an estimated effect size of $d = 0.84$. In the simultaneous fits (see Table 6), the between- vs. within-subjects moderator was again the only significant moderator.

We next performed meta-analyses on the 46 (out of 56) effect sizes in the test format category for which initial test performance data were available (Table 6). When initial test performance was tested as a single candidate moderator, it was not significant on its own ($p = .12$). However, when initial test performance was simultaneously fit to the data along with the aforementioned seven candidate moderators, it was significant along with the between- vs. within-subjects design, response congruency, and multiple-choice vs. not multiple-choice on the initial test moderators (ps $\leq .048$). Between-paper heterogeneity was reduced to zero in that final model (τ^2 = zero), whereas within paper heterogeneity remained high ($\omega^2 = 0.11$). The findings for initial test performance and response congruency were also consistent with patterns observed in the overall meta-analyses.

(Table 6 around here)

Analyses adjusted for publication bias. In PEESE analyses to the entire test format

category dataset with moderators fitted (Table 7; results of analyses with no moderators also shown), *sv* was not a significant moderator of effect size ($p = .29$). Moreover, the intercept effect size ($d = 0.39$) was similar to that in the corresponding random-effects analysis ($d = 0.49$). The between- vs. within-subjects design moderator effect also remained positive but was no longer significant in the model ($p = .37$).

In PEESE analyses with moderators limited to data for which initial test performance was reported, *sv* was a significant moderator of effect size ($p = .025$), suggesting publication bias. However, the moderators of response congruency, initial test accuracy, and multiple choice vs. not on the initial test (that were identified in the random-effects analysis) all remained potent. The between- vs. within-subjects design moderator was again not significant. Because that moderator did not emerge from the overall analysis and was not identified in any other category analyses, we infer that it has a weak effect on transfer at best (in contrast to its apparent influence on testing effects, as detailed in Rowland, 2014), and do not analyze it further.

(Table 7 around here)

In sensitivity analyses to the entire test format category dataset, the intercept estimate was generally unaffected by the different scenarios of publication bias (Table 8). The sole exception was the case of severe one-tailed publication bias, in which the intercept was reduced by more than half but remained positive. Similarly, sensitivity analyses limited to data for which initial test performance was reported revealed only modest changes to effect size estimates for both the moderators and the intercept, again excepting the case of severe one-tailed publication bias.

(Table 8 around here)

Stimulus-response rearrangement. A category-level meta-analysis on 33 effect sizes from all 10 studies in the stimulus-response rearrangement category (including four sub-categories of stimulus types: paired associates, triple associates, multi-term facts, and term-

definition facts) with no candidate moderators fitted yielded a weighted mean effect size of d = 0.22 (a small effect), p = .066, C.I. [─0.019, 0.45], suggesting weak transfer at best. Between-paper heterogeneity was again observed, however (τ^2 = 0.071, ω^2 = zero). Inspection of the forest plot in panel b of Figure 2 reveals a more nuanced pattern: positive transfer for paired associates and minimal transfer elsewhere (as observed by Pan & Rickard, 2017).

Single and simultaneous moderator fits. We tested the following three candidate moderators using random-effects meta-analysis: number of training phase item repetitions, retention interval, and paired associates vs. non-paired associates. The values of other aforementioned candidate moderators exhibited minimal variability across experiments and were not analyzed. In the single moderator fits (Table 6), only the paired associates vs. non-paired associates moderator was significant. For non-paired associate stimuli, the weighted effect size was negligible, d = 0.063, p = .29, C.I. [─0.078, 0.20]; for paired associates, the estimated effect size increased by d = 0.66, p = .0031, C.I. [0.46, 0.86], yielding a large estimated effect size of d = 0.72. Although the p-value for paired associates should be treated with some caution given insufficient degrees of freedom for that category at the paper level, its small value, in combination with consistent results over studies (see Figure 2, panel b), gives us high subjective confidence in the conclusion of strong transfer for that case. When all three candidate moderators were evaluated simultaneously (see Table 6), only the paired associates vs. non-paired associates moderator survived. Heterogeneity was reduced to near zero in that final model (τ^2 = 0.0020, ω^2 = zero).

Analyses adjusted for publication bias. In PEESE analyses to the stimulus-response rearrangement category that included the paired associates vs. non-paired associates moderator (Table 7), *sv* was marginally significant (p = .064). The intercept estimate of roughly zero for non-paired associate stimuli corresponds closely to the random-effects analysis, as does the

estimated effect size increment of d = 0.59 for paired associate stimuli. In the sensitivity analyses, the effect size estimate for the paired associate vs. non-paired associate moderator, as well as the intercept, exhibited only modest fluctuations under the four scenarios of publication bias (Table 8).

Untested materials seen during initial study. A category-level meta-analysis on 17 effect sizes from the 12 studies in the untested materials category (with no candidate moderators fitted) yielded a weighted mean effect size of d = 0.16, p = .20, C.I. [−0.10, 0.43]. Thus, there is no compelling evidence of transfer for this category. There was however relatively high between-paper heterogeneity (τ^2 = 0.11), but zero within-paper heterogeneity. The heterogeneity between papers is evident upon inspection of the forest plot in panel b of Figure 2.

Single and simultaneous moderator fits. We tested the following five candidate moderators individually using random-effects meta-analysis: between- vs. within-subjects design, number of training phase item repetitions, correct answer feedback, retention interval, and elaborated retrieval practice. In the single moderator fits (Table 6), only the elaborated retrieval practice moderator was significant, d = 0.37, p = .032, C.I. [0.041, 0.70]; the use of such strategies increased the estimated effect size from d = 0.0028 to 0.37. Between-paper heterogeneity was modestly reduced in that moderator fit (τ^2 = 0.077, ω^2 = zero). The results for that moderator reflect corresponding findings in the overall meta-analyses, as well as prior suggestions in the literature for this transfer category (e.g., Balch, 1998; Chan et al., 2006; Little, 2011; Hinze et al., 2013; see also Chan, 2009; 2010; Little & Bjork, 2015; Little, Bjork, Bjork, & Angello, 2012). When all five candidate moderator were evaluated simultaneously (see Table 6), again only the elaborated retrieval practice moderator was significant.

Analyses adjusted for publication bias. In PEESE analyses that included the elaborated retrieval practice moderator (Table 7), *sv* was a significant moderator (p = .028), suggesting

publication bias in the untested materials category. The adjusted intercept estimate was negative (d = ─0.32), suggesting that transfer to untested materials is worse in the testing than non-testing reexposure control conditions (in contrast, in the random-effects analyses, the intercept estimate was essentially zero). The estimated effect size increase for elaborated retrieval practice (d = 0.34) was significant and, in contrast with nearly every other PEESE analysis performed, larger than in the random-effects model. However, given the negative intercept estimate of similar magnitude, there is by this analysis no evidence of positive transfer in this category relative to a non-testing reexposure control, even when elaborated retrieval practice is present. It thus appears that that testing with elaborated retrieval practice yields learning in this category that is equivalent to that in non-testing re-exposure conditions such as restudy.

The sensitivity analyses (Table 8) converge with the conclusions based on the PEESE analysis. The effect size estimate for elaborated retrieval practice was minimally affected by the different scenarios, and the intercept was negative in all cases.

Application and inference questions. A random-effects analysis on 41 effect sizes from the 17 papers in the application and inference category (including three sub-categories: application questions, inference questions, or both) with no candidate moderators fitted yielded a weighted mean effect size of d = 0.32, p = .0013, C.I. [0.085, 0.56], indicating overall positive transfer. However, there was substantial between-paper (τ^2 = 0.11) but zero within-paper heterogeneity.

Single and simultaneous moderator fits. We tested the following five candidate moderators using random-effects meta-analysis, the values of which varied sufficiently over experiments: between- vs. within-subjects design, retention interval, correct answer feedback, and elaborated retrieval practice. No sub-category comparisons (e.g., application vs. inference questions) were performed due to the limited number of papers in each sub-category; moreover,

visual inspection of the category-level forest plot suggests that the variation in effect sizes in this category is not attributable to sub-category. In the single moderator fits (Table 6), the number of training phase item repetitions, retention interval, and elaborated retrieval practice moderators were significant (ps ≤ .029). An increase in training repetitions or retention interval was associated with improved transfer (estimated increases of d = 0.33, p = .016, C.I. [0.11, 0.55] and 0.0033, C.I. [0.0010, 0.0054] for each additional repetition and added hour, respectively), as was the use of elaborated retrieval practice (estimated increase of d = 0.35, p = .029, C.I. [0.046, 0.66]). When all five moderators were evaluated simultaneously (see Table 6), the correct answer feedback moderator was significant (p = .011), and the number of training phase repetitions and elaborated retrieval practice moderators were marginally significant (ps ≤ .063). However, the p-value for correct answer feedback and the number of training phase repetitions should be treated with caution due to insufficient degrees of freedom. Correct answer feedback was associated with less transfer (estimated decrease of d = ─0.49, C.I. [─0.75, ─0.24]), yielding an estimated effect size of d = ─0.049, whereas elaborated retrieval practice was associated with the reverse (estimated increase of d = 0.26, C.I. [─0.019, 0.54]), yielding an estimated effect size of d = 0.70. Between-paper heterogeneity was substantially reduced in that final model (τ^2 = 0.021, ω^2 = zero).

Analyses adjusted for publication bias. In the PEESE analyses of the application and inference category dataset with moderators fitted (Table 7), all three of the moderating variables that were identified in the random-effects analysis survived with similar effect size estimates. Although there was a trend toward publication bias as indicated by the estimated coefficient for *sv*, it did not reach statistical significance (p = .11). In sensitivity analyses to the same dataset (Table 8), the effect size estimates for all moderators and the intercept were minimally affected by the different scenarios of publication bias.

Problem-solving skills. A category-level meta-analysis on 17 effect sizes from the nine papers that comprise the problem-solving skills category (including two sub-categories: medical diagnosis and treatment and worked examples) with no candidate moderators fitted yielded a weighted mean effect size of $d = 0.29$, $p = .10$, C.I. [−0.078, 0.65], indicating weak transfer. However, as is confirmed upon inspection of the forest plot in Figure 2, panel d, as well as by the substantial between-paper heterogeneity measure ($\tau^2 = 0.13$, $\omega^2 =$ zero), there is a sizeable difference between the results for the medical diagnosis and treatment sub-category and those for the worked examples sub-category.

Single and simultaneous moderator fits. The difference between the two sub-categories of problem-solving skills was confirmed using random-effects meta-analysis by fitting worked examples vs. medical diagnosis and treatment as a single moderator (Table 6), $d = 0.59$, $p = .028$, C.I. [0.093, 1.09]. The use of problem types which involve medical diagnosis and treatment increased the estimated effect size from $d = 0.045$ to $d = 0.59$. Between-paper heterogeneity was substantially reduced in that moderator fit ($\tau^2 = 0.047$, $\omega^2 =$ zero). None of four other candidate moderators fitted in this case (between- vs. within-subjects design, correct answer feedback, number of training phase item repetitions, and retention interval) approached statistical significance. When all five candidate moderators were fitted simultaneously (see Table 6), again only the worked examples vs. medical diagnosis and treatment moderator was significant.

Analyses adjusted for publication bias. In PEESE analyses to the problem-solving category dataset which included the worked examples vs. medical diagnosis and treatment moderator (Table 7), the results were a close match to that of the random-effects analysis, with no trend suggesting publication bias ($p = .99$). In sensitivity analyses to the problem-solving skills dataset (Table 8), the effect size estimate for the worked examples vs. medical diagnosis and treatment moderator was largely unaffected except for the case of severe one-tailed

publication bias, and the intercept was near zero in all cases.

Mediator and related word cues. A meta-analysis on 27 effect sizes from the five papers that comprise the mediator and related word cues category yielded a weighted mean effect size of $d = 0.61$ (a medium-large effect), $p = .018$, C.I. [0.25, 0.97], indicating positive transfer (although the *p*-value should be treated with caution due to insufficient degrees of freedom at the paper level). Although inferentially valid moderator fits were not possible due to insufficient data, we were able to estimate separate weighted mean effect sizes for the two cue types: mediator cues ($d = 0.76$); related cues ($d = 0.47$). That numerical difference is consistent with paper-level statistical results (e.g., Carpenter, 2011) and is evident upon inspection of panel d of Figure 2.

Analyses adjusted for publication bias. In PEESE analyses to the mediator and related word cues dataset (no moderators were analyzed in this category; see Table 7), *sv* was marginally significant ($p = .070$), suggesting possible publication bias, and the adjusted intercept ($d = 0.36$) was reduced relative to that of the random-effects analysis. In sensitivity analyses of the same dataset, the intercept was minimally affected by the different scenarios of publication bias.

Supplementary Analyses of Testing vs. Transfer Effects

In the test-enhanced learning literature, differing predictions about the extent to which transfer effects may be smaller or larger than testing effects have been made (e.g., Carpenter & DeLosh, 2006; McDaniel, Anderson, et al., 2007; Rohrer et al., 2010). It has also been an open question in the literature as to whether testing and transfer effects are correlated with one another. To address both issues, we plotted testing and transfer effect size data from all 81 experiments in our sample that assessed both effects on the final test (approximately 40% of effect sizes in our dataset, encompassing five categories, had such information). The mean testing effect size in our dataset, $d = 0.68$, is roughly comparable to that observed in prior meta-

analyses (e.g., weighted mean effect sizes of g = 0.70 in Adescope et al., 2017 and g = 0.50 in Rowland, 2014), although it is important to note that those studies involved different sets of articles identified using a different set of selection criteria.

Results for testing vs. transfer effects are shown in Figure 3. Most data points are below the diagonal (dotted line) that corresponds to equivalent effect sizes (56 of 81 cases; binomial test: p = .00070). Hence, on average, transfer effects are smaller than testing effects, a pattern that may generally be the case in this literature. The only exception was a non-significant trend toward larger transfer effects in the test format category.

(Figure 3 around here)

An additional and unexpected result evident in Figure 3 is that, in the overall data set, testing and transfer effect sizes are at best weakly correlated. However, that result may mask contrasting patterns within category type. We will return to that topic in the Discussion.

Discussion

The foregoing meta-analyses investigated the wealth of accumulated research on transfer of test-enhanced learning. In aggregate across that literature, there is substantial evidence that testing can yield positive transfer relative to non-testing reexposure control conditions such as restudy and rereading. When considering the fact that transfer is often notoriously difficult to achieve (Gick & Holyoak, 1987; Haskell, 2001; Singley & Anderson, 1989), and that its very existence has been debated (e.g., Barnett & Ceci, 2002; Detterman, 1993; Singley & Anderson, 1989), that finding is in itself notable. However, in some categories and sub-categories, weak, null, or even negative transfer was observed, and particularly in analyses that adjust for publication bias. Overall across the literature, positive transfer of test-enhanced learning appears to be strongly conditional on key aspects of performance and experiment design – a finding that informs the principles of transfer that we propose next.

A Three-Factor Framework for Transfer of Test-Enhanced Learning

Drawing upon the three major moderators uncovered in the overall meta-analyses, we propose a *three-factor framework* for transfer of test-enhanced learning, according to which transfer manifests as a function of whether there is *response congruency* between the initial and final tests, whether *elaborated retrieval practice* (i.e., broad encoding methods and/or elaborative feedback) is employed during training, and whether the level of *initial test performance* is high or low. The predictions of that framework for the overall dataset wherein initial test accuracy was reported are depicted graphically in Figure 4, panel a, where the estimated moderator effects are from the random-effects meta-analysis (PEESE moderator results, shown in panel b, were roughly equivalent, although the no-moderator intercept was generally reduced). In the random-effects analysis, the between-paper heterogeneity was reduced nearly three-fold, from $\tau^2 = 0.077$ to $\tau^2 = 0.028$. It thus appears that the major determinants of transfer across papers are captured by that model. In contrast, within-paper heterogeneity was virtually unchanged after the model was fitted ($\omega^2 = 0.061$). However, that effect appears to be exclusively driven by the test format category. In the previously reported random-effects analysis limited to that category, within-paper heterogeneity was very large ($\omega^2 = 0.11$), even after moderators were identified; correspondingly, when that category was removed from the overall analyses post hoc, ω^2 dropped to zero.[4]

(Figure 4 around here)

Under the conditions of no response congruency, no elaborated retrieval practice, and low

[4] Inspection of details of the articles in the test format category appears to explain the high value of ω^2. Multiple papers examined variations within a single test format across multiple experiments (e.g., varying the difficulty of the cues presented, as in Halamish & Bjork, 2011; Rowland & DeLosh, 2016), as well as other differences in training phase design (e.g., mixed vs. pure lists in Rowland et al., 2014; different sequences of study and test trials in Jacoby et al., 2010). Those variables could not be addressed in the present meta-analyses.

initial test proportion correct (i.e., by subtracting 0.19, the lowest value of that initial test proportion correct in the dataset, from every effect size, yielding a new intercept), the estimated transfer effect size in the random-effects analysis is near zero, with a confidence interval which at its upper extreme yields only a small negative transfer effect, d = ─0.053, C.I. [─0.22, 0.12]. In contrast, with both response congruency and the use of elaborated retrieval practice present in that same model, and with the initial test accuracy intercept set to the maximum value in the dataset (0.98), the predicted effect size is d = 0.90, C.I. [0.71, 1.08]. These two contrasting cases illustrate the descriptive power of the three-factor framework.

Similar results were obtained in the PEESE analyses, although with notably reduced effect sizes. Under the conditions of no response congruency, no elaborated retrieval practice, and low initial test performance, the estimated transfer effect size is negative (d = ─0.21). At the opposite extreme, with response congruency and elaborated retrieval practice present, and initial test accuracy at the maximum value in the dataset, the predicted effect size is d = 0.58.

In the next sections we consider each of the moderating factors in the three-factor transfer framework in further detail.

Response congruency. If strong response congruency holds in a given study, then there is an increased likelihood of positive transfer. The effect of response congruency is generally consistent with identical elements and other similarity-based models of transfer (e.g., Thorndike, 1906; for related discussions see Hamaker, 1986; Morris et al., 1977; Roediger & Blaxton, 1987; Tulving, 1970, 1984). However, its precise definition differs from those used in other accounts of similarity in which response congruency is not an explicit focus, and which refer more broadly (and often far less precisely) to semantic or other processing similarities between the training and transfer contexts (e.g., Anderson, 1993; Bruce, 1933; Haskell, 2001; Morris et al., 1977), or which propose different mechanisms (e.g., Healy, Wohldmann, & Bourne, 2005). One

prominent exception is Wylie (1919), who proposed that transfer is determined by the "objective similarity" between the two responses (cf. Osgood, 1949).

Indeed, the response congruency effect appears in some cases to be in direct contradiction of the predictions of other similarity-based frameworks. This is illustrated by the following case of stimulus-response rearrangement involving facts (from Pan, Gopal, et al., 2015): the initial test involved the question, *"Thomas Jefferson purchased Louisiana from WHOM?"* and a subsequent transfer test involved the question, *"Thomas Jefferson purchased WHAT from Spain?"* At both surface and semantic levels, there appears to be strong similarity between those questions, and thus most similarity-based transfer frameworks would appear to predict positive transfer, whereas in this case none was observed.

The effect of response congruency as a predictor of transfer can be further appreciated by inspection of differences in the correlations between testing and transfer effects for categories that did or did not have that property (see Figure 3). Response congruency held for all studies in the mediator and related word cues categories and for most cases in the test format category, and across those two cases there were indications of a positive correlation between testing and transfer effects. In contrast, in the three other categories wherein response congruency generally did not hold, no such correlation was evident.

Mechanisms of positive transfer via response congruency. There are two theoretical reasons to expect that response congruency may facilitate transfer. First, if a correct response on the final test was also retrieved or provided as correct answer feedback on the initial test, then that response may be more available (i.e., more easily accessible) on the final test than would otherwise be the case (Bjork & Bjork, 1992; Carrier & Pashler, 1991; Vaughn & Rawson, 2014). A clear case in which that effect may be at play is the presentation of related cues on the final test for the same responses that were retrieved on the initial test (e.g., Rawson et al., 2015).

Related cues are by design only weakly related to the correct final test response, and thus may be unlikely to facilitate retrieval of the correct response by themselves. However, if the correct response was made more available by the initial test, then the joint factors of that increased availability and the weak association with the related cue may boost final test performance above that of the non-testing reexposure control condition.

Second, in many of the cases in which response congruency holds, it is also the case that all or part of the stimulus-to-response pathway that was established on the initial test (at least for correct trials) can be reinstated to support retrieval of the correct answer on the final test. Four such cases of *stimulus-to-response pathway reinstatement* are depicted in Figure 5 (all panels). Consider first the case of transfer from an initial multiple-choice test to a final cued recall test (i.e., across test formats), in which the correct responses on the initial and final tests are the same, and excepting a format change, the initial and final test cues are as well (e.g., Meyer & Logan, 2013). As shown in panel a, the originally learned stimulus-to-response pathway can be reinstated (i.e., by the transfer stimulus, which differs from the initial test stimulus only by a change in the presented test format) to retrieve the correct response on the final test, thus yielding direct transfer of test-enhanced learning. The same can apply for the reverse case (i.e., cued recall to multiple-choice). Another plausible case involves mediator word cues on the final test (e.g., the aforementioned example of *father* as a mediator for *mother-child*). Two scenarios of stimulus-to-response pathway reinstatement for that case have been advanced in the literature (Cho et al., 2017; Coppens et al, 2016). In the first, depicted in panel b, the mediator word cue accesses a mediator-to-target pathway (e.g. *father-child*) that is presumed to have been formed or strengthened on the initial test. In the second, depicted in panel c, the mediator word cue (e.g., *father*) activates the initial test word cue (e.g., *mother*), thus reinstating the full stimulus-to-response pathway (e.g., *father* elicits *mother*, which then elicits the target, *child*) that was

strengthened on the initial test.

(Figure 5 around here)

Complete or partial stimulus-to-response pathway reinstatement is also likely in studies involving medical diagnosis and treatment. In the former case (e.g., Kromann et al., 2009), which corresponds to panel a of Figure 5, the patient demographics and/or symptoms on the initial and final tests may differ slightly, but the overall scenarios and procedures (i.e., a cardiac resuscitation checklist) are nearly or completely identical. In that case, the cues on the final test likely can support reinstatement of the full stimulus-to-response pathway. In the latter case (e.g., Larsen et al., 2013a; Larsen et al., 2013b), which corresponds to panel d, the stimuli on the transfer test partially overlap with (or are a subset of) those on the initial test. For example, after training to recognize and treat different neurological conditions, subjects are presented with patient scenarios which correspond to those conditions on a final test. In that case, partial stimulus-to-response pathway reinstatement is likely to occur (i.e., subjects are able to link the cues that are presented on the transfer test to previously trained symptoms and procedures).

Elaborated retrieval practice. The use of elaborated retrieval practice increases the likelihood of positive transfer. We have suggested that elaborated retrieval practice takes two distinct and often non-overlapping forms: broad encoding methods and elaborative feedback.

Broad encoding methods. The use of broad retrieval instructions, discrimination instructions, explanatory recall, and the combination of high and low order questions can be efficacious at yielding transfer. Indeed, all nine effect sizes across five meta-analyzed studies in four categories that involved broad encoding methods yielded statistically significant positive transfer effects (Chan et al., 2006; Hinze et al., 2013; Larsen et al., 2013b; Little, 2011; McDaniel, Bugg, et al., 2015). Broad encoding methods may be especially critical for yielding transfer to untested materials seen during initial study, as such methods likely result in additional

processing of those materials (Anderson & Biddle, 1975; Little, 2011). For instance, in Chan et al. (2006; Experiment 3), subjects in the broad retrieval condition were instructed to think of all related information while generating answers to initial test questions; positive transfer to untested materials was subsequently observed in that condition, but not in other conditions which lacked such instructions. One candidate underlying mechanism for transfer via broad encoding methods is greater cognitive processing of initially studied materials (for discussions see Carpenter & DeLosh, 2006; Chan et al., 2006; Frase, 1968). That added processing may stem from reactivation, or reminding, of memories formed during initial study that are not limited to the correct answer but may also involve other aspects of the studied materials (for further discussion of the effects of reminding, see Jacoby, Wahlheim, & Kelley, 2015; Tullis, Benjamin, & Ross, 2014). Alternatively, more indirect processes (e.g., discrimination between answer choices as discussed by Little, 2011; improved construction of mental models as suggested by Hinze et al., 2013) may underlie transfer via broad encoding methods.

It is important to reemphasize here that only four types of training methods qualified as broad encoding methods in the meta-analyses, and together these studies only comprised a small proportion of the literature. The definition of broad encoding methods is open to expansion or reinterpretation, and the effectiveness of other techniques that might qualify remains to be explored. Indeed, some training techniques that might have been expected to yield similar effects evidently do not; for example, in the case of stimulus-response rearrangement, even the retrieval of two terms or words per fact or word triplet across separate training trials has been shown to yield no positive transfer to untested responses (e.g., Pan, Wong, et al., Experiment 2), a result that is, however, fully consistent with the response congruency factor. In another example, the use of multiple rephrased initial test questions does not necessarily yield better transfer to application and inference questions than does repeat presentations of identical

questions (Butler, 2010). It may be the case that broad encoding methods have to foster or improve the retrieval of associations between multiple pieces of information acquired during initial study (and not just isolated pieces of information) in order to improve transfer.

Elaborative feedback. Post-retrieval processing of elaborative feedback (i.e., restudy of all to-be-learned information, explanatory feedback, or extended and detailed feedback) can generate positive transfer. In fact, positive transfer was observed for 36 of 40 cases from 18 studies, spanning across five transfer categories, which featured elaborative feedback. That finding supports prior claims on the *indirect effects* of testing (i.e., activities associated with but not directly involved in the act of testing itself, such as improved restudy; for further discussion see Roediger & Karpicke, 2006) for transfer performance (e.g., Balch, 1998; McDaniel, Howard, et al., 2009; McDaniel & Little, in press; Nguyen & McDaniel, 2016; Pan, Gopal, et al., 2015; Pan, Wong, et al., 2016; Pan & Rickard, 2017; van Eersel et al., 2016; and others).

By contrast, simple correct answer feedback was not associated with improved transfer performance in any of the meta-analyses. Thus, it appears that, to reliably enhance transfer, feedback must include more than just the correct response. This stands in contrast with the results for the larger test-enhanced learning literature, in which correct answer feedback by itself is often sufficient to substantially boost testing effect magnitude (Rowland, 2014).

Initial test performance. The retrieval success rate on the initial test also substantially predicts the magnitude of transfer in this literature. That finding is consistent with a positive correlation in the dataset between initial test performance and transfer effect size (which is shown in Figure 6). A candidate account of that effect is that high accuracy on the initial test reflects not only better memory for the target information that is tested, but also more complete memory for other aspects of the study event, including any inferences or other thoughts that occurred during the initial study phase that may be relevant for the transfer task. By this

account, when initial test accuracy is high, those other memory aspects are relatively likely to have been retrieved along with the correct answer on the initial test, yielding (via test-enhanced learning) relatively high probability of retrieval of those memory aspects on a final transfer test (i.e., positive transfer relative to a non-testing reexposure control condition). In contrast, low initial test accuracy likely correlates with partial, or piecewise retrieval of the study event (i.e., when overall retrieval performance for the target information is poor, those memories are presumably also less complete). When initial test performance is low, sometimes the retrieved pieces of the memory include retrieval of the correct answer, but most often it does not. In either case, the probability that aspects of the initial study event that may be relevant for transfer will be retrieved on the initial test is expected to be lower than would be the case when initial test accuracy is high, leading to poor transfer. This account, along with other candidate accounts of the effect of initial test performance on transfer, warrants further investigation.

(Figure 6 around here)

The three-factor transfer framework and category-level results. In each instance wherein a moderator in the three-factor transfer framework could be analyzed via simultaneous moderator fits at the category level, it emerged as statistically significant in both the random-effects and PEESE analyses. There were four such cases: response congruency and initial test performance in the test format category, and elaborated retrieval practice in the untested materials and application and inference categories (in other categories, such analyses were precluded by the absence of within-category moderator variability). These results reinforce the importance of those three factors as general determinants of transfer.

Category-Specific Moderators

There were six cases across four categories in which a moderator was identified for only one category. We briefly address those findings in the following section.

Test format. The multiple-choice vs. not on the initial test moderator emerged as significant for this category in both the random-effects and PEESE analyses, with reduced transfer for the multiple-choice case. That result is consistent with prior hypotheses in the literature (e.g., Duchastel, 1981; Foos & Fisher, 1991; McDaniel, Anderson, et al., 2007; McDaniel, McDermott, et al., 2007; McDermott et al., 2014; Rickard & Pan, 2017; Wenger, Thompson, & Bartling, 1980).

Stimulus-response rearrangement. The paired associates vs. non-paired associates moderator was significant in this category in both the random-effects and PEESE analyses. That finding is consistent with prior observations on the different transfer properties of stimuli with two vs. three or more elements (e.g., Pan & Rickard, 2017). The basis for that contrasting effect for superficially similar materials is investigated in the context of the dual memory model in Rickard and Pan (2018).

Application and inference questions. Correct answer feedback (present or absent) emerged as a significant moderator of transfer in the negative direction in this category. By contrast, elaborated retrieval practice, which in this category most commonly took the form of elaborative feedback, was a significant moderator in the opposite direction (as occurred in the overall meta-analyses). It thus appears that feedback which involves more than just the correct answer (such as post-retrieval restudy of entire text passages or extended or explanatory feedback; for examples see Blunt & Karpicke, 2014, Experiment 2; Eglington & Kang, 2016; Karpicke & Blunt, 2011, Experiment 1; McDaniel, Howard, et al., 2009, Experiment 2; and Zhou et al., 2013, Experiment 2) is needed for feedback to improve transfer in this category. When just the correct answer is provided, transfer may not manifest or may even be negative (for examples see Agarwal, 2011, Experiment 1; Nguyen, Gouravajhala, & McDaniel, 2016; Tran, Rohrer, & Pashler, 2014; Wooldridge et al., 2014, Experiment 1).

Two other moderators, number of training phase repetitions and retention interval, were also significant in this category only. It is not clear, however, why both of these moderators would be predictive only for the case of application and inference questions and not in the overall analyses (as substantial variability in both moderators is present across multiple categories). In light of the expectation that Type I errors may occur across the large number of tested moderators at the category level, results for these moderators should be regarded as tentative.

Problem-solving skills. The significance of the worked examples vs. medical diagnosis and treatment moderator in this category reflects the starkly different results for its two constituent sub-categories. It reflects sub-category differences in (a) subject materials, (b) training methods (i.e., problem worksheets without feedback vs. a range of testing and feedback methods), (c) non-testing reexposure control conditions (i.e., worked example study vs. text restudy), (d) settings (i.e., laboratory vs. clinical), and (e) problem-solving procedures. Due to the limited data currently available, it is not possible to fully disentangle these content, design, and procedural differences (which may conflate content with processes; despite that possibility, the two sub-categories accurately characterize this category as it currently exists). It is however plausible that testing might yield different degrees of transfer across different problem types. For new results that address the general absence of feedback in the worked examples sub-category and other implementation factors, see Yeo and Fazio, (in press).

Publication Bias

PEESE analyses of the overall dataset indicated at least moderate publication bias in the transfer of test-enhanced learning literature. Nevertheless, in all cases the moderator variables that were identified in the random-effects analysis remained significant, with modestly reduced effect estimates (see Figure 7 for all effect sizes plotted against *sv*, in effect a funnel plot turned

sideways, along with PEESE moderator estimates). The net result was that, when none of the moderators took values that enhanced transfer, the predicted transfer effect was null or negative.

At the category level, the PEESE analyses with moderators fitted produced statistically significant evidence for publication bias only in the test format and untested materials categories. That result may appear to contradict the substantial bias detected in the overall analysis. On closer examination, however, there was a trend toward publication bias for every category except for problem-solving skills. Further, the estimated slopes of the *sv* parameter for those categories (in PEESE analyses with moderators listed in Table 7) are 3.19, 6.23, 4.35, 5.68, and 1.94, which compares favorably to the *sv* estimate of 3.86 in the overall PEESE analyses. That result is consistent with the conclusion that publication bias is pervasive across most categories in this literature but was not detected within most categories due to the smaller sample size and lower statistical power. Finally, both overall and across categories, the sensitivity analyses indicated that a moderate level of publication bias would decrease intercept effect sizes by about the magnitude predicted by the PEESE analyses.

(Figure 7 around here)

Limitations of the Current Work

As in any regression or meta-regression analysis, causal inference regarding statistically significant moderators should be tentative. Nevertheless, it is notable that each of the major moderators in the three-factor framework lends itself to plausible and relatively straightforward causal interpretation. Moreover, in most cases they are consistent with prior empirical or theoretical work. First, although response congruency has not specifically been considered in the current test-enhanced learning literature as a factor that may moderate transfer (indeed, it appears not to have been prominently considered for a wide range of transfer contexts since Osgood, 1949), it is sensible and generally consistent with principles of learning that it would. Moreover,

there is independent evidence that making a response increases its subsequent retrieval availability (e.g., Estes, 1979; for review see Vaughn & Rawson, 2014), and the aforementioned scenarios of stimulus-to-response pathway reinstatement constitute a case of "near" transfer that is consistent with multiple theoretical frameworks of memory (e.g., Healy et al., 2005). Second, the multiple forms of elaborated retrieval practice identified in this review have all been hypothesized to yield transfer in the prior literature, with some prior experiment-level support (e.g., Chan et al., 2006; Hinze et al., 2013; McDaniel, Bugg, et al., 2015). Finally, a higher level of initial test performance may reflect better and broader learning of target materials (as we elaborated earlier), again yielding a higher likelihood of positive transfer.

The meta-analyses are also potentially limited by the simplifying assumptions that were made during the calculation of between- and within-subjects effect sizes, as previously detailed. Accordingly, although between- vs. within-subjects design did not emerge as a significant predictor in the overall meta-analyses and did not survive the simultaneous model fits of the PEESE analyses, our conclusions regarding that moderator should be regarded as tentative. Moreover, with regards to the moderators identified in this review, there are always the possibilities that some candidate moderators were not detected due to insufficient power, and that one or more important moderators have not yet been hypothesized.

Category-specific moderators. For some category-level analyses (e.g., mediator and related word cues, as well as problem-solving skills) there were only a small number of studies and effect sizes. Thus, although the estimates of aggregate effect size at the category level should generally be trustworthy, inferences about candidate moderating factors that could be analyzed at that level should be made with caution.

Implications for Theories of Test-Enhanced Learning and Transfer

As detailed at the outset of this review, few theoretical accounts of test-enhanced learning

directly address transfer. Moreover, the wide range of qualitatively different learning and transfer contexts compounds the challenge of adapting existing theoretical accounts to address them all. However, the constituent components of the three-factor transfer framework – which identify important encoding, retrieval, and design factors for transfer in this literature – can be integrated within existing theoretical perspectives. First, the effect of response congruency is the most readily accounted for by theories of test-enhanced learning that are framed, at least in part, in terms of associative memory, such that the concepts of response availability and stimulus-to-response pathway reinstatement can be readily incorporated. Examples include the elaborative retrieval (Carpenter & DeLosh, 2006), mediator effectiveness (Pyc & Rawson, 2009), and dual memory (Rickard & Pan, 2017) theories. Second, with regard to elaborated retrieval practice, the case of positive transfer where broad encoding methods are used is potentially consistent with accounts which specifically reference a test-induced process of spreading activation (e.g., Carpenter, 2009; Chan et al., 2006; Pyc & Rawson, 2009). Additionally, the effectiveness of elaborative feedback for transfer is consistent with qualitative accounts of test-enhanced learning which focus on its indirect effects (e.g., Arnold & McDermott, 2013; Balch, 1998; McDaniel, Howard, et al., 2009; McDaniel and Little, in press; Nguyen et al., 2016; Nguyen & McDaniel, 2016; Pan, Gopal, et al., 2015; Pan & Rickard, 2017; van Eersel et al., 2016). Uncovering the precise mechanistic basis for the effectiveness of such training techniques is an important goal for future work.

With regards to the broader transfer literature, the three-factor framework readily connects to the transfer frameworks of Perkins and Salomon (1994) and Barnett and Ceci (2002). In relation to the former framework, the response congruency effect can be construed as “low” circumstances of transfer, whereas elaborated retrieval practice may yield more abstract learning and thus constitute “high” circumstances of transfer (therefore the three-factor transfer

framework incorporates perspectives from both the identical elements and related similarity-based models of transfer, as well as the general principle and other abstractionist models). In relation to the latter framework, both response congruency and elaborated retrieval practice may enhance learners' ability to recall and execute prior learning on the transfer test. The three-factor transfer framework also highlights how the structure of the training and transfer tasks, as well as encoding and retrieval factors, can moderate the magnitude of transfer (as suggested by Gick & Holyoak, 1987). Finally, as previously discussed, our findings for the response congruency moderator provide support for prior theories of response similarity and transfer (Wylie, 1933).

Potential Educational and Practical Implications

We propose that this review's findings can be distilled into the following four educationally relevant principles of transfer. We caution the reader, however, that these principles are derived from research in predominantly laboratory settings, and that the studies to date cover only a sample of the variety of transfer contexts of educational interest.

1. *Transfer is likeliest when the answers on the initial and final tests are the same, and less likely when they are not*. By definition, transfer tests involve questions that are different in some way from those that were previously encountered. However, if the correct answers to those new questions are the same (or nearly so) as the correct answers to questions that were used during training (e.g., whereas an initial test question asks for a term given a definition, and the transfer test asks for the same term given a real-world example), then transfer is more likely.
2. *Transfer increases when initial tests involve retrieving information broadly (broad encoding).* Transfer is more likely when initial tests involve discriminating between different concepts, constructing an explanation, or recalling a specific concept using several questions that address different levels of knowledge (i.e., high and low order

questions as described in Bloom's taxonomy).

3. *Transfer increases when post-testing involves restudy of materials or other forms of elaborative feedback is provided.* Combining practice testing with restudy of all to-be-learned materials, feedback containing explanations, or feedback that is presented for extended review increases the likelihood of transfer.
4. *Transfer increases with higher accuracy on the initial test.* The better one performs on the initial test, the more likely transfer will occur. This candidate principle suggests that manipulations that enhance initial test accuracy – such as achievement of a relatively high level of learning prior to testing – may enhance transfer.

Although the foregoing meta-analyses suggest potentially promising avenues, as well as limitations, on the effective use of testing to foster transfer, more research is needed into how best to implement practice testing to induce transfer in authentic educational and other training contexts.

Directions for Future Research

For further insights into transfer of test-enhanced learning, new empirical research is needed. Promising avenues include:

1. *New or under-explored transfer contexts* – currently, there are only one or a few published studies regarding category learning and classification (e.g., Baghdady et al., 2014; Jacoby et al., 2010), visuospatial skills (Carpenter & Kelly, 2012; Rohrer et al., 2010; cf. Kelly, Carpenter, & Sjolund, 2015), and function learning (Kang et al., 2011). There are also a number of other prominent transfer contexts that remain almost entirely unresearched in the test-enhanced learning literature (e.g., analogical transfer and other types of abstract learning), as well as existing transfer categories that remain relatively underpopulated in terms of number of studies (e.g., problem-solving skills; mediator and

related word cues). Finally, the literature is currently dominated by more "near" than "far" transfer studies (e.g., very few studies address transfer across knowledge domains).

2. *Transfer in authentic educational contexts* – with several prominent exceptions (e.g., Agarwal, Bain, & Chamberlain, 2012; Balch, 1998; Bjork, Little, & Storm, 1994; McConnell, St-Onge, & Young, 2014; McDaniel, Anderson, et al., 2007; McDaniel, Wildman, et al., 2012), the vast majority of research in this literature has occurred in laboratory settings. However, an ultimate aim of test-enhanced learning research is to develop practice testing into a portable and effective real world learning technique. Thus, future studies should further investigate testing and transfer in actual classroom and other learning environments, using methods that approach the level of control afforded by laboratory studies to the extent practicable. Such research might be more likely to incorporate more instances of "far" transfer than have been examined in the literature to date (e.g., transfer across multiple different contexts).
3. *The theoretical and mechanistic basis for transfer of test-enhanced learning* – in tandem with further research on theoretical mechanisms of the testing effect, more research is needed into the cognitive processes that yield transfer of test-enhanced learning, as well as the circumstances under which that learning may or may not transfer (i.e., from both within and outside the three-factor transfer framework).
4. *Further investigations of elaborated retrieval practice* – both broad encoding methods and elaborative feedback were meta-analyzed for the first time in this review. However, the number of studies featuring elaborated retrieval practice remains limited, the implementation of such methods varies substantially between studies (e.g., broad retrieval vs. discrimination instructions), and the circumstances under which their effectiveness is optimized remains to be determined.

5. *The ecological validity and role of stimulus materials* – in several categories, the choice of to-be-learned materials dramatically affected or appeared to affect transfer results (e.g., paired associates vs. non-paired associates in the stimulus-response rearrangement category; medical diagnoses vs. worked examples in the problem-solving category; also potentially highly semantically related vs. not semantically related questions in the untested materials category). Further research into such effects is needed. Also, some researchers (e.g., Butler, 2010; Chan et al., 2006; Wooldridge et al., 2014) have suggested that the stimulus materials used in some studies appear to be relatively "contrived" and have questionable ecological validity; further research is also needed to explore that possibility.
6. *Other moderating factors* – due to insufficient data, potentially relevant factors that could not be investigated in the foregoing meta-analyses include the similarity of cues between the initial and transfer tests, the role of prior knowledge, the use of hints and reminders at the transfer test (e.g., to alert learners to the need to transfer knowledge across knowledge domains, as occurred in Butler, 2010; Experiment 3), and test expectancy manipulations e.g., Hinze & Rapp, 2014; Nguyen & McDaniel, 2016). There may also be other, as-yet unidentified moderating factors.
7. *Other paradigms involving test-enhanced learning* – implementations of testing that have shown promise at yielding transfer include successive relearning (also called mastery learning; e.g., Rawson, Dunlosky, & Sciartelli, 2013) and interpolated testing (e.g., Wissman, Rawson & Pyc, 2011). Both involve different training methods than the standard test-enhanced learning paradigm.

Conclusions

In the foregoing meta-analyses, we established that test-enhanced learning can yield

transfer performance that is often *substantially better than* that in non-testing reexposure control conditions such as restudy or rereading. Among the major transfer contexts investigated in the literature to date, testing generally yields positive transfer across test formats and to application and inference questions, mediator and related word cues, and problems involving medical diagnoses; it often yields numerically weak transfer, or in some cases possibly negative transfer, to stimulus-response rearranged items, untested materials seen during initial study, and problems involving worked examples (although there are a number of prominent exceptions). Publication bias appears to be moderate in this literature, reducing the magnitude of, but not eliminating, positive transfer. Important moderators of transfer include response congruency, elaborated retrieval practice, and the level of initial test performance. Together, these moderating factors form the basis of a new three-factor transfer framework that appears to accommodate the majority of results in this literature and provides insights into optimizing transfer in applied settings.

References

*References marked with an asterisk indicate studies included in the meta-analyses.

Abbott, E. E. (1909). On the analysis of the factor of recall in the learning process. *The Psychological Review: Monograph Supplements, 11*(1), 159-177. doi:http://dx.doi.org/10.1037/h0093018

Adesope, O. O., Trevisan, D. A., & Sundararajan, N. (2017). Rethinking the use of tests: A meta-analysis of practice testing (vol 87, pg 659, 2017). *Review of Educational Research, 87*(3), 1-1.

*Agarwal, P. K. (2011). *Examining the relationship between fact learning and higher order learning via retrieval practice* (Order No. AAI3468823).

Agarwal, P. K., Bain, P. M., & Chamberlain, R. W. (2012). The value of applied research: Retrieval practice improves classroom learning and recommendations from a teacher, a principal, and a scientist. *Educational Psychology Review, 24*(3), 437-448. doi:http://dx.doi.org/10.1007/s10648-012-9210-2

Agarwal, P. K., Finley, J. R., Rose, N. S., & Roediger, H. L. (2016). Benefits from retrieval practice are greater for students with lower working memory capacity. *Memory* (advance online publication), doi:http://dx.doi.org/10.1080/09658211.2016.1220579

Allport, G. W. (1937). *Personality: A psychological interpretation* Holt: Oxford.

Anderson, J. R. (1996). ACT: A simple theory of complex cognition. *American Psychologist, 51*(4), 355-365. doi:http://dx.doi.org/10.1037/0003-066X.51.4.355

Anderson, L., & Krathwohl, D. A. (2001) *Taxonomy for learning, teaching and assessing: a revision of Bloom's taxonomy of educational objectives.* New York: Longman.

Anderson, R. C., Kulhavy, R. W., & Andre, T. (1972). Conditions under which feedback facilitates learning from programmed lessons. *Journal of Educational Psychology, 63*(3),

186-188. doi:http://dx.doi.org/10.1037/h0032653

Anderson, R., & Biddle, W. (1975). On asking people questions about what they are reading. In G. H. Bower (Ed.), *The Psychology of Learning and Motivation*, (pp. 90-132). New York: Academic Press.

Arnold, K. M., & McDermott, K. B. (2013). Test-potentiated learning: Distinguishing between direct and indirect effects of tests. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 39*(3), 940-945. doi:http://dx.doi.org/10.1037/a0029199

APA Publications and Communications Board Working Group on Journal Article Reporting Standards. (2008). Reporting standards for research in psychology: Why do we need them? What might they be? *American Psychologist*, 63, 839–851. doi:http://dx.doi.org/10.1037/0003-066X.63.9.839

*Avci, G. (2011). *Transfer of the testing effect: Just how powerful is it?* (Order No. AAI3464205).

Baghdady, M., Carnahan, H., Lam, E. W. N., & Woods, N. N. (2014). Test-enhanced learning and its effect on comprehension and diagnostic accuracy. *Medical Education,* 48(2), 181-188. doi:http://dx.doi.org/10.1111/medu.12302

*Balch, W. R. (1998). Practice versus review exams and final exam performance. *Teaching of Psychology,* 25(3), 181-185.

Barnett, S. M., & Ceci, S. J. (2002). When and where do we apply what we learn?: A taxonomy for far transfer. *Psychological Bulletin,* 128(4), 612-637. doi:http://dx.doi.org/10.1037/0033-2909.128.4.612

Benjamin, A. S., & Pashler, H. (2015). The value of standardized testing: a perspective from cognitive psychology. *Policy Insights from the Behavioral and Brain Sciences*, 2(1), 13-23.

*Bies-Hernandez, N. (2014). *Examining the testing effect using the dual-process signal detection model* (Order No. AAI3590123).

Bjork, E. L., Little, J. L., & Storm, B. C. (2014). Multiple-choice testing as a desirable difficulty in the classroom. *Journal of Applied Research in Memory and Cognition,* 3(3), 165-170. doi:http://dx.doi.org/10.1016/j.jarmac.2014.03.002

Bjork, R. A., & Bjork, E. L. (1992). A new theory of disuse and an old theory of stimulus fluctuation. *Essays in honor of William K. Estes, vol. 1: From learning theory to connectionist theory; vol. 2: From learning processes to cognitive processes.* (pp. 35-67). Hillsdale: Lawrence Erlbaum Associates, Inc.

Bjork, R.A. (1975). Retrieval as a memory modifier: An interpretation of negative recency and related phenomena. In R.L. Solso (Ed.), *Information processing and cognition: The Loyola Symposium* (pp. 123–144). Hillsdale: Erlbaum.

Bloom, B. S. (1956). *Taxonomy of educational objectives: The classification of educational goals (1st ed.)* Harlow: Longman Group.

Bloom, B. S. (1984). *Taxonomy of educational objectives*. Boston: Allyn and Bacon.

*Blunt, J. R., & Karpicke, J. D. (2014). Learning with retrieval-based concept mapping. *Journal of Educational Psychology, 106*(3), 849-858. doi:http://dx.doi.org/10.1037/a0035934

Borenstein, M., Hedges, L. V., Higgins, J., & Rothstein, H. R. (2010). A basic introduction to fixed-effect and random-effects models for meta-analysis. *Research Synthesis Methods, 1,* 97–111. doi:http://dx.doi.org/10.1002/jrsm.12

Bourne, L. E., Jr., & Healy, A. F. (2014). *Train your mind for peak performance: A science-based approach for achieving your goals* Washington, DC: American Psychological Association. doi:http://dx.doi.org/10.1037/14319-000

Brookhart, S. M. (2015). Making the Most of Multiple Choice. *Educational Leadership*, *73*(1),

36-39.

Brooks, L. W., & Dansereau, D. F. (1987). Transfer of information: An instructional perspective. *Transfer of learning: Contemporary research and applications.* (pp. 121-150) San Diego: Academic Press.

Brown, P. C., Roediger, H. L., & McDaniel, M. A. (2014). *Make it stick: The science of successful learning*. Cambridge: Belknap Press.

Bruce, R. W. (1933). Conditions of transfer of training. *Journal of Experimental Psychology, 16*(3), 343-361. doi:http://dx.doi.org/10.1037/h0074550

Bujang, M. A., & Baharum, N. (2017). Guidelines of the minimum sample size requirements for Kappa agreement test. *Epidemiology, Biostatistics and Public Health*, *14*(2).

*Butler, A. C. (2010). Repeated testing produces superior transfer of learning relative to repeated studying. *Journal of Experimental Psychology: Learning, Memory, and Cognition,* 36(5), 1118-1133. doi:http://dx.doi.org/10.1037/a0019902

*Butler, A. C., & Roediger, H. L. (2007). Testing improves long-term retention in a simulated classroom setting. *European Journal of Cognitive Psychology,* 19(4-5), 514-527. doi:http://dx.doi.org/10.1080/09541440701326097

Butler, A. C., Godbole, N., & Marsh, E. J. (2013). Explanation feedback is better than correct answer feedback for promoting transfer of learning. *Journal of Educational Psychology,* 105(2), 290-298. doi:http://dx.doi.org/10.1037/a0031026

Butler, A. C., Karpicke, J. D., & Roediger, H. L.. (2007). The effect of type and timing of feedback on learning from multiple-choice tests. *Journal of Experimental Psychology: Applied, 13*(4), 273-281. doi:http://dx.doi.org/10.1037/1076-898X.13.4.273

Carey, B. (2013). Frequent tests can enhance college learning, study finds. *The New York Times.* Available at: http://www.nytimes.com/2013/11/21/education/frequent-tests-can-enhance-

college-learning-study-finds.html

*Carpenter, S. K. (2009). Cue strength as a moderator of the testing effect: The benefits of elaborative retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition,* 35(6), 1563-1569. doi:http://dx.doi.org/10.1037/a0017021

*Carpenter, S. K. (2011). Semantic information activated during retrieval contributes to later retention: Support for the mediator effectiveness hypothesis of the testing effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition,* 37(6), 1547-1552. doi:http://dx.doi.org/10.1037/a0024140

Carpenter, S. K. (2012). Testing enhances the transfer of learning. *Current Directions in Psychological Science, 21*(5), 279-283. doi:http://dx.doi.org/10.1177/0963721412452728

Carpenter, S. K., & DeLosh, E. L. (2006). Impoverished cue support enhances subsequent retention: Support for the elaborative retrieval explanation of the testing effect. *Memory & Cognition,* 34(2), 268-276. doi:http://dx.doi.org/10.3758/BF03193405

Carpenter, S. K., & Kelly, J. W. (2012). Tests enhance retention and transfer of spatial learning. *Psychonomic Bulletin & Review,* 19(3), 443-448. doi:http://dx.doi.org/10.3758/s13423-012-0221-2

Carpenter, S. K., Lund, T. J. S., Coffman, C. R., Armstrong, P. I., Lamm, M. H., & Reason, R. D. (2016). A classroom study on the relationship between student achievement and retrieval-enhanced learning. *Educational Psychology Review, 28*(2), 353-375. doi:http://dx.doi.org/10.1007/s10648-015-9311-9

*Carpenter, S. K., & Pashler, H. (2007). Testing beyond words: Using tests to enhance visuospatial map learning. *Psychonomic Bulletin & Review, 14*(3), 474-478. doi:http://dx.doi.org/10.3758/BF03194092

*Carpenter, S. K., Pashler, H., & Vul, E. (2006). What types of learning are enhanced by a cued

recall test? *Psychonomic Bulletin & Review,* 13(5), 826-830. doi:http://dx.doi.org/10.3758/BF03194004

Carpenter, S. K., Pashler, H., Wixted, J. T., & Vul, E. (2008). The effects of tests on learning and forgetting. *Memory & Cognition, 36*(2), 438-448. doi:http://dx.doi.org/10.3758/MC.36.2.438

Carrier, M., & Pashler, H. (1992). The influence of retrieval on retention. *Memory & Cognition, 20*(6), 633-642.

Chan, J. C. K. (2009). When does retrieval induce forgetting and when does it induce facilitation? implications for retrieval inhibition, testing effect, and text processing. *Journal of Memory and Language*, 61(2), 153-170. doi:http://dx.doi.org/10.1016/j.jml.2009.04.004

Chan, J. C. K. (2010). Long-term effects of testing on the recall of nontested materials. *Memory,* 18(1), 49-57. doi:http://dx.doi.org/10.1080/09658210903405737

*Chan, J. C. K., McDermott, K. B., & Roediger, H. L. (2006). Retrieval-induced facilitation: Initially nontested material can benefit from prior testing of related material. *Journal of Experimental Psychology: General,* 135(4), 553-571. doi:http://dx.doi.org/10.1037/0096-3445.135.4.553

*Cheng, C. K. (2014). *Effect of multiple-choice testing on memory retention -- cue-target symmetry* (Order No. AAI3666589).

*Cho, K. W., Neely, J. H., Brennan, M. K., Vitrano, D., & Crocco, S. (2017). Does testing increase spontaneous mediation in learning semantically related paired associates? *Journal of Experimental Psychology: Learning, Memory, and Cognition 43*(11).

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

Collins, A. M., & Loftus, E. F. (1975). A spreading-activation theory of semantic processing. *Psychological Review, 82*(6), 407-428. doi:http://dx.doi.org/10.1037/0033-295X.82.6.407

*Coppens, L. C., Verkoeijen, P. P. J. L., Bouwmeester, S., & Rikers, R. M. J. P. (2016). The testing effect for mediator final test cues and related final test cues in online and laboratory experiments. *BMC Psychology, 4*, 14.

Cormier, S. M, & Hagman, J. D. (Eds). *Transfer of learning: Contemporary research and applications* (1987). San Diego: Academic Press.

Cranney, J., Ahn, M., McKinnon, R., Morris, S., & Watts, K. (2009). The testing effect, collaborative learning, and retrieval-induced facilitation in a classroom setting. *European Journal of Cognitive Psychology, 21*(6), 919-940. doi:http://dx.doi.org/10.1080/09541440802413505

*Darabi, A. A., Nelson, D. W., & Palanki, S. (2007). Acquisition of troubleshooting skills in a computer simulation: Worked example vs. conventional problem solving instructional strategies. *Computers in Human Behavior, 23*(4), 1809-1819. doi:http://dx.doi.org/10.1016/j.chb.2005.11.001

Delaney, P. F., Verkoeijen, P. P. J. L., & Spirgel, A. (2010). Spacing and testing effects: A deeply critical, lengthy, and at times discursive review of the literature. In B. H. Ross (Ed.), *The psychology of learning and motivation: Advances in research and theory (vol. 53);* (pp. 63-147). San Diego: Academic Press. doi:http://dx.doi.org/10.1016/S0079-7421(10)53003-2

Dempster, F. N. (1996). Distributing and managing the conditions of encoding and practice. *Memory.* (pp. 317-344). San Diego: Academic Press. doi:http://dx.doi.org/10.1016/B978-012102570-0/50011-2

Detterman, D. K. (1993). The case for the prosecution: Transfer as an epiphenomenon. *Transfer*

on trial: Intelligence, cognition, and instruction. (pp. 1-24) Westport: Ablex Publishing.

Druckman, D., & Bjork, R. A. (1994). *Learning, remembering, believing: Enhancing human performance.* Washington, DC: National Academy Press.

Duchastel, P. C. (1981). Retention of prose following testing with different types of tests. *Contemporary Educational Psychology,* 6(3), 217-226.

Duchastel, P. C., & Nungester, R. J. (1982). Testing effects measured with alternate test forms. *The Journal of Educational Research*, 75(5), 309-313.

Dudai, Y. (2007). Transfer: its transfer into neurobiology. In H. L. Roediger, Y. Dudai & S. M. Fitzpatrick (Eds.), *Science of memory: Concepts,* New York: Oxford University Press.

Dunlap, W. P., Cortina, J. M., Vaslow, J. B., & Burke, M. J. (1996). Meta-analysis of experiments with matched groups or repeated measures designs. *Psychological Methods, 1*(2), 170-177. doi:http://dx.doi.org/10.1037/1082-989X.1.2.170

Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., & Willingham, D. T. (2013). Improving students' learning with effective learning techniques: Promising directions from cognitive and educational psychology. *Psychological Science in the Public Interest, 14*(1), 4-58. doi:http://dx.doi.org/10.1177/1529100612453266

Ebbinghaus, H. (1885). *Über das Gedächtnis: Untersuchungen zur experimentellen Psychologie*. Duncker & Humblot.

*Eglington, L. G., & Kang, S. H. K. (2016). Retrieval practice benefits deductive inference. *Educational Psychology Review,* doi:http://dx.doi.org/10.1007/s10648-016-9386-y

Ellis, H. C. (1965). *The transfer of learning*. Oxford: Macmillan.

Estes, W. K. (1979). Role of response availability in the effects of cued-recall tests on memory. *Journal of Experimental Psychology: Human Learning and Memory, 5*(6), 567-573. doi:http://dx.doi.org/10.1037/0278-7393.5.6.567

Fiorella, L., & Mayer, R. E. (2015). Eight ways to promote generative learning. *Educational Psychology Review,* doi:http://dx.doi.org/10.1007/s10648-015-9348-9

Foos, P. W., & Fisher, R. P. (1988). Using tests as learning opportunities. *Journal of Educational Psychology,* 80(2), 179-183. doi:http://dx.doi.org/10.1037/0022-0663.80.2.179

Frase, L. T. (1968). Effect of question location, pacing, and mode upon retention of prose material. *Journal of Educational Psychology, 59*(4), 244-249. doi:http://dx.doi.org/10.1037/h0025947

Gasparinatou, A., & Grigoriadou, M. (2013). Exploring the effect of background knowledge and text cohesion on learning from texts in computer science. *Educational Psychology, 33*(6), 645-670. doi:http://dx.doi.org/10.1080/01443410.2013.790309

George, T., & Wiley, J. (2016, November). *Going the distance: The effects of testing on analogical transfer.* Poster presented at the Psychonomic Society Annual Meeting, Boston, MA.

Gick, M. L., & Holyoak, K. J. (1980). Analogical problem solving. *Cognitive Psychology,* 12(3), 306-355.

Gick, M. L., & Holyoak, K. J. (1987). The cognitive basis of knowledge transfer. *Transfer of learning: Contemporary research and applications.* (pp. 9-46). San Diego: Academic Press.

Glass, G. V., McGaw, B., Smith, M. L. (1981). Meta-analysis in social research. Newbury Park: Sage Publications.

Glover, J. A. (1989). The "testing" phenomenon: Not gone but nearly forgotten. *Journal of Educational Psychology*, 81(3), 392-399. doi:http://dx.doi.org/10.1037/0022-0663.81.3.392

Goode, M. K., Geraci, L., & Roediger, H. L. (2008). Superiority of variable to repeated practice

in transfer on anagram solution. *Psychonomic Bulletin & Review*, 15(3), 662-666. doi:http://dx.doi.org/10.3758/PBR.15.3.662

*Halamish, V., & Bjork, R. A. (2011). When does testing enhance retention? A distribution-based interpretation of retrieval as a memory modifier. *Journal of Experimental Psychology: Learning, Memory, and Cognition,* 37(4), 801-812. doi:http://dx.doi.org/10.1037/a0023219

Hamaker, C. (1986). The effects of adjunct questions on prose learning. *Review of Educational Research, 56*(2), 212-242. doi:http://dx.doi.org/10.2307/1170376

Hanawalt, N. G., & Tarr, A. G. (1961). The effect of recall upon recognition. *Journal of Experimental Psychology,* 62(4), 361-367. doi:http://dx.doi.org/10.1037/h0041917

Haskell, R. E. (2001). *Transfer of learning: Cognition, instruction, and reasoning*. San Diego: Academic Press.

Healy, A. F. (2007). Transfer: specificity and generality. In H. L. Roediger, Y. Dudai & S. M. Fitzpatrick (Eds.), *Science of memory: Concepts,* New York: Oxford University Press.

Healy, A. F., Wohldmann, E. L., & Bourne, L. E., Jr. (2005). The procedural reinstatement principle: Studies on training, retention, and transfer. In A. F. Healy (Ed.), *Experimental cognitive psychology and its applications; experimental cognitive psychology and its applications* (pp. 59-71, Chapter xxii, 265 Pages) American Psychological Association, Washington, DC. doi:http://dx.doi.org/10.1037/10895-005

Hedges, L. V., Tipton, E., & Johnson, M. C. (2010). Robust variance estimation in meta-regression with dependent effect size estimates. *Research Synthesis Methods, 1*(1), 39–65. doi:http://dx.doi.org/10.1002/jrsm.5

*Hinze, S. R., & Rapp, D. N. (2014). Retrieval (sometimes) enhances learning: Performance pressure reduces the benefits of retrieval practice. *Applied Cognitive Psychology, 28*(4),

597-606. doi:http://dx.doi.org/10.1002/acp.3032

*Hinze, S. R., & Wiley, J. (2011). Testing the limits of testing effects using completion tests. *Memory,* 19(3), 290-304. doi:http://dx.doi.org/10.1080/09658211.2011.560121

*Hinze, S. R., Wiley, J., & Pellegrino, J. W. (2013). The importance of constructive comprehension processes in learning from tests. *Journal of Memory and Language,* 69(2), 151-164. doi:http://dx.doi.org/10.1016/j.jml.2013.03.002

Hogan, R. M., & Kintsch, W. (1971). Differential effects of study and test trials on long-term recognition and recall. *Journal of Verbal Learning & Verbal Behavior*, 10(5), 562-567.

*Huff, M. J., Balota, D. A., & Hutchison, K. A. (2016). The costs and benefits of testing and guessing on recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 42*(10), 1559-1572. doi:http://dx.doi.org/10.1037/xlm0000269

Huff, M. J., Coane, J. H., Hutchison, K. A., Grasser, E. B., & Blais, J. E. (2012). Interpolated task effects on direct and mediated false recognition: Effects of initial recall, recognition, and the ironic effect of guessing. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 38*(6), 1720-1730. doi:http://dx.doi.org/10.1037/a0028476

*Jacoby, L. L., Wahlheim, C. N., & Coane, J. H. (2010). Test-enhanced learning of natural concepts: Effects on recognition memory, classification, and metacognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition,* 36(6), 1441-1451. doi:http://dx.doi.org/10.1037/a0020636

Jacoby, L. L., Wahlheim, C. N., & Kelley, C. M. (2015). Memory consequences of looking back to notice change: Retroactive and proactive facilitation. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 41*(5), 1282-1297. doi:http://dx.doi.org/10.1037/xlm0000123

James, W. (1890). *The principles of psychology, vol I.* New York: Henry Holt and Co.

doi:http://dx.doi.org/10.1037/10538-000

Jensen, J. L., McDaniel, M. A., Woodard, S. M., & Kummer, T. A. (2014). Teaching to the test ... or testing to teach: Exams requiring higher order thinking skills encourage greater conceptual understanding. *Educational Psychology Review,* 26(2), 307-329. doi:http://dx.doi.org/10.1007/s10648-013-9248-9

*Johnson, C. I., & Mayer, R. E. (2009). A testing effect with multimedia learning. *Journal of Educational Psychology,* 101(3), 621-629. doi:http://dx.doi.org/10.1037/a0015183

Jones, A. C., Wardlow, L., Pan, S. C., Zepeda, C., Heyman, G. D., Dunlosky, J., and Rickard, T. C. (2015). Beyond the rainbow: retrieval practice leads to better learning than does rainbow writing. *Educational Psychology Review 27*(3). doi: 10.1007/s10648-015-9330-6

Judd, C. H. (1908). The relation of special training to general intelligence. *Educational Review, 36,* 28–42.

Kang, S. H. K., McDaniel, M. A., & Pashler, H. (2011). Effects of testing on learning of functions. *Psychonomic Bulletin & Review,* 18(5), 998-1005. doi:http://dx.doi.org/10.3758/s13423-011-0113-x

*Kang, S. H. K., McDermott, K. B., & Roediger, H. L. (2007). Test format and corrective feedback modify the effect of testing on long-term retention. *European Journal of Cognitive Psychology,* 19(4-5), 528-558. doi:http://dx.doi.org/10.1080/09541440601056620

Karpicke, J. D. (2012). Retrieval-based learning: Active retrieval promotes meaningful learning. *Current Directions in Psychological Science*, *21*(3), 157-163.

Karpicke, J. D., & Aue, W. R. (2015). The testing effect is alive and well with complex materials. *Educational Psychology Review*, 27(2), 317-326. doi:http://dx.doi.org/10.1007/s10648-015-9309-3

*Karpicke, J. D., & Blunt, J. R. (2011). Retrieval practice produces more learning than elaborate studying with concept mapping. *Science,* 331(6018), 772-775.

*Karpicke, J. D., Blunt, J. R., Smith, M. A., & Karpicke, S. S. (2014). Retrieval-based learning: The need for guided retrieval in elementary school children. *Journal of Applied Research in Memory and Cognition,* 3(3), 198-206. doi:http://dx.doi.org/10.1016/j.jml.2009.11.010

Karpicke, J. D., Lehman, M., & Aue, W. R. (2014). Retrieval-based learning: An episodic context account. *The psychology of learning and motivation (vol 61).* (pp. 237-284). San Diego: Elsevier Academic Press.

Kelly, E. L. (1967). Transfer of training: an analytic study. In B. P. Kosimar & C. J. B. MacMillan (Eds.), *Psychological concepts in education* (p. 50). Chicago: Rand McNally.

Kelly, J. W., Carpenter, S. K., & Sjolund, L. A. (2015). Retrieval enhances route knowledge acquisition, but only when movement errors are prevented. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 41*(5), 1540-1547. doi:http://dx.doi.org/10.1037/a0038685

Kornell, N., Bjork, R. A., & Garcia, M. A. (2011). Why tests appear to prevent forgetting: A distribution-based bifurcation model. *Journal of Memory and Language, 65*(2), 85-97. doi:http://dx.doi.org/10.1016/j.jml.2011.04.002

*Kromann, C. B., Bohnstedt, C., Jensen, M. L., & Ringsted, C. (2010). The testing effect on skills learning might last 6 months. *Advances in Health Sciences Education, 15*(3), 395-401. doi:http://dx.doi.org/10.1007/s10459-009-9207-x

*Kromann, C. B., Jensen, M. L., & Ringsted, C. (2011). Test-enhanced learning may be a gender-related phenomenon explained by changes in cortisol level. *Medical Education, 45*(2), 192-199. doi:http://dx.doi.org/10.1111/j.1365-2923.2010.03790.x

*Kromann, C. B., Jensen, M. L., & Ringsted, C. (2009). The effect of testing on skills learning.

Medical Education, 43(1), 21-27. doi:http://dx.doi.org/10.1111/j.1365-2923.2008.03245.x

Kulhavy, R. W., & Anderson, R. C. (1972). Delay-retention effect with multiple-choice tests. *Journal of Educational Psychology, 63*(5), 505-512. doi:http://dx.doi.org/10.1037/h0033243

Kulhavy, R. W., & Stock, W. A. (1989). Feedback in written instruction: The place of response certitude. *Educational Psychology Review, 1*(4), 279-308. doi:http://dx.doi.org/10.1007/BF01320096

Kuo, T., & Hirshman, E. (1996). Investigations of the testing effect. *The American Journal of Psychology, 109*(3), 451-464.

Lahey, J. (2014). Students should be tested more, not less. *The Atlantic.* Available at: http://www.theatlantic.com/education/archive/2014/01/students-should-be-tested-more-not-less/283195/

Lakens, D., Hilgard, J., & Staaks, J. (2016). On the reproducibility of meta-analyses: Six practical recommendations. *BMC Psychology, 4*, 10.

*LaPorte, R. E., & Voss, J. F. (1975). Retention of prose materials as a function of postacquisition testing. *Journal of Educational Psychology,* 67(2), 259-266. doi:http://dx.doi.org/10.1037/h0076933

*Larsen, D. P., Butler, A. C., & Roediger, H. L. (2013a). Comparative effects of test-enhanced learning and self-explanation on long-term retention. *Medical Education, 47*(7), 674-682. doi:http://dx.doi.org/10.1111/medu.12141

*Larsen, D. P., Butler, A. C., Lawson, A. L., & Roediger, H. L. (2013b). The importance of seeing the patient: Test-enhanced learning with standardized patients and written tests improves clinical application of knowledge. *Advances in Health Sciences Education,*

18(3), 409-425. doi:http://dx.doi.org/10.1007/s10459-012-9379-7

Laubscher, N. F. (1960). Normalizing the noncentral t and F distributions. *The Annals of Mathematical Statistics*, *31*(4), 1105-1112.

*Leahy, W., Hanham, J., & Sweller, J. (2015). High element interactivity information during problem solving may lead to failure to obtain the testing effect. *Educational Psychology Review,* 27(2), 291-304.

*Lechuga, M. T., Ortega-Tudela, J., & Gómez-Ariza, C. J. (2015). Further evidence that concept mapping is not better than repeated retrieval as a tool for learning from texts. *Learning and Instruction, 40*, 61-68. doi:http://dx.doi.org/10.1016/j.learninstruc.2015.08.002

*Little, J. L. (2011). *Optimizing multiple-choice tests as learning events* (Order No. AAI3493389).

Little, J. L., & Bjork, E. L. (2015). Optimizing multiple-choice tests as tools for learning. *Memory & Cognition,* 43(1), 14-26. doi:http://dx.doi.org/10.3758/s13421-014-0452-8

Little, J. L., Bjork, E. L., Bjork, R. A., & Angello, G. (2012). Multiple-choice tests exonerated, at least of some charges: Fostering test-induced learning and avoiding test-induced forgetting. *Psychological Science,* 23(11), 1337-1344. doi:http://dx.doi.org/10.1177/0956797612443370

Mandler, G., & Rabinowitz, J. C. (1981). Appearance and reality: Does a recognition test really improve subsequent recall and recognition? *Journal of Experimental Psychology: Human Learning and Memory,* 7(2), 79-90. doi:http://dx.doi.org/10.1037/0278-7393.7.2.79

Marzano, R. J.; Pickering, D.; Pollock, J. E. (2006). *Classroom instruction that works: research-based strategies for increasing student achievement*. ASCD.

Mayer, R. E. (2009). *Multimedia learning*. New York: Cambridge University Press.:http://dx.doi.org/10.1017/CBO9780511811678

McConnell, M. M., St-Onge, C., & Young, M. E. (2015). The benefits of testing for learning on later performance. *Advances in Health Sciences Education,* 20(2), 305-320. doi:http://dx.doi.org/10.1007/s10459-014-9529-1

McDaniel, M. A. (2007). Transfer: rediscovering a central concept. In H. L. Roediger, Y. Dudai & S. M. Fitzpatrick (Eds.), *Science of memory: Concepts,* New York: Oxford University Press.

McDaniel, M. A., & Fisher, R. P. (1991). Tests and test feedback as learning sources. *Contemporary Educational Psychology,* 16(2), 192-201.

McDaniel, M. A., & Little, J. L. (in press). Multiple-choice and short-answer quizzing on equal footing in the classroom: potential indirect effects of testing. In J. Dunlosky & K. Rawson (Eds.), *Handbook of Cognition and Education*, Cambridge University Press.

McDaniel, M. A., & Masson, M. E. (1985). Altering memory representations through retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition,* 11(2), 371-385. doi:http://dx.doi.org/10.1037/0278-7393.11.2.371

*McDaniel, M. A., Anderson, J. L., Derbish, M. H., & Morrisette, N. (2007). Testing the testing effect in the classroom. *European Journal of Cognitive Psychology,* 19(4-5), 494-513. doi:http://dx.doi.org/10.1080/09541440701326154

*McDaniel, M. A., Bugg, J. M., Liu, Y., & Brick, J. (2015). When does the test-study-test sequence optimize learning and retention? *Journal of Experimental Psychology: Applied,* 21(4), 370-382. doi:http://dx.doi.org/10.1037/xap0000063

*McDaniel, M. A., Howard, D. C., & Einstein, G. O. (2009). The read-recite-review study strategy: Effective and portable. *Psychological Science,* 20(4), 516-522. doi:http://dx.doi.org/10.1111/j.1467-9280.2009.02325.x

McDaniel, M. A., Roediger, H. L., & McDermott, K. B. (2007). Generalizing test-enhanced

learning from the laboratory to the classroom. *Psychonomic Bulletin & Review, 14*(2), 200-206. doi:http://dx.doi.org/10.3758/BF03194052

McDaniel, M. A., Thomas, R. C., Agarwal, P. K., McDermott, K. B., & Roediger, H. L. (2013). Quizzing in middle-school science: Successful transfer performance on classroom exams. *Applied Cognitive Psychology*, 27(3), 360-372. doi:http://dx.doi.org/10.1002/acp.2914

*McDaniel, M. A., Wildman, K. M., & Anderson, J. L. (2012). Using quizzes to enhance summative-assessment performance in a web-based class: An experimental study. *Journal of Applied Research in Memory and Cognition,* 1(1), 18-26. doi:http://dx.doi.org/10.1016/j.jarmac.2011.10.001

McDermott, K. B., Agarwal, P. K., D'Antonio, L., Roediger, H. L., & McDaniel, M. A. (2014). Both multiple-choice and short-answer quizzes enhance later exam performance in middle and high school classes. *Journal of Experimental Psychology: Applied*, 20(1), 3-21. doi:http://dx.doi.org/10.1037/xap0000004

McGeoch, J. A. (1942). *The psychology of human learning: An introduction.* Longmans: Oxford.

McNamara, D. S., & Kintsch, W. (1996). Learning from texts: Effects of prior knowledge and text coherence. *Discourse Processes, 22*(3), 247-288. doi:http://dx.doi.org/10.1080/01638539609544975

Mestre, J. P. (2005). *Transfer of learning from a modern multidisciplinary perspective.* IAP.

*Meyer, A. N. D., & Logan, J. M. (2013). Taking the testing effect beyond the college freshman: Benefits for lifelong learning. *Psychology and Aging,* 28(1), 142-147. doi:http://dx.doi.org/10.1037/a0030890

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & The PRISMA Group (2009). Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. *PLoS Med 6*(7): e1000097. doi:http://dx.doi.org/10.1371/journal.pmed1000097

Morris, C. D., Bransford, J. D., & Franks, J. J. (1977). Levels of processing versus transfer appropriate processing. *Journal of Verbal Learning & Verbal Behavior,* 16(5), 519-533.

Morris, S. B., & DeShon, R. P. (2002). Combining effect size estimates in meta-analysis with repeated measures and independent-groups designs. *Psychological Methods, 7*(1), 105-125. doi:http://dx.doi.org/10.1037/1082-989X.7.1.105

Mozer, M. C., Howe, M., & Pashler, H. (2004). Using testing to enhance learning: A comparison of two hypotheses. In K. Forbus, D. Gentner, & T. Regier (Eds.), *Proceedings of the Twenty-Sixth Annual Conference of the Cognitive Science Society* (pp. 975–980). Mahwah: Erlbaum.

*Mulligan, N. W., & Peterson, D. J. (2015). Negative and positive testing effects in terms of item-specific and relational information. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 41*(3), 859-871. doi:http://dx.doi.org/10.1037/xlm0000056

*Nguyen, K., & McDaniel, M. A. (2016). The JOIs of text comprehension: Supplementing retrieval practice to enhance inference performance. *Journal of Experimental Psychology: Applied, 22*(1), 59-71. doi:http://dx.doi.org/10.1037/xap0000066

*Nguyen, K., Gouravajhala, R., & McDaniel, M. A. (2016). Can testing enhance transfer of learning and is it retrieval driven? Unpublished manuscript.

*Nungester, R. J., & Duchastel, P. C. (1982). Testing versus review: Effects on retention. *Journal of Educational Psychology, 74*(1), 18-22. doi:http://dx.doi.org/10.1037/0022-0663.74.1.18

Osgood, C. E. (1949). The similarity paradox in human learning: A resolution. *Psychological Review, 56*(3), 132-143. doi:http://dx.doi.org/10.1037/h0057488

*Pan, S. C., Gopal, A., & Rickard, T. C. (2015). Testing with feedback yields potent, but piecewise, learning of history and biology facts. *Journal of Educational Psychology*

107(4). doi:http://dx.doi.org/ 10.1037/edu0000074

*Pan, S. C., Hutter, S., D'Andrea, D., Unwalla, D., & Rickard, T. C. Investigations of learning and transfer following elaborated retrieval practice on scientific concepts. Unpublished manuscript.

Pan, S. C., Pashler, H., Potter, Z. E., & Rickard, T. C. (2015). Testing enhances learning across a range of episodic memory abilities. *Journal of Memory and Language 83*, 53-61. doi:http://dx.doi.org/10.1016/j.jml.2015.04.001

Pan, S. C., & Rickard, T. C. (2015). Sleep and motor memory: is there room for consolidation? *Psychological Bulletin* 141(4). doi:http://dx.doi.org/10.1037/bul0000009

*Pan, S. C., & Rickard, T. C. (2017). Does retrieval practice enhance learning and transfer for term-definition facts? *Journal of Experimental Psychology: Applied, 23*(3). doi:http://dx.doi.org/10.1037/xap0000124

Pan, S. C., Rubin, B. R., & Rickard, T. C. (2015). Does testing with feedback improve adult spelling skills relative to copying and reading? *Journal of Experimental Psychology: Applied 21*(4). doi:http://dx.doi.org/ 10.1037/xap0000062

*Pan, S. C., Wong, C. M., Potter, Z. E., Mejia, J., & Rickard, T. C. (2015). Does test-enhanced learning transfer for triple associates? *Memory & Cognition 44*(1). doi:http://dx.doi.org/ 10.3758/s13421-015-0547-x

Pashler, H., Bain, P., Bottge, B., Graesser, A., Koedinger, K., McDaniel, M., & Metcalfe, J. (2007). *Organizing instruction and study to improve student learning* (NCER 2007–2004). Washington, DC: National Center for Education Research, Institute of Education Sciences, U.S. Department of Education. Available from: http://ncer.ed.gov.

Pashler, H., Cepeda, N. J., Wixted, J. T., & Rohrer, D. (2005). When does feedback facilitate learning of words? *Journal of Experimental Psychology: Learning, Memory, and*

Cognition, 31(1), 3-8. doi:http://dx.doi.org/10.1037/0278-7393.31.1.3

Pashler, H., Rohrer, D., Cepeda, N. J., & Carpenter, S. K. (2007). Enhancing learning and retarding forgetting: Choices and consequences. *Psychonomic Bulletin & Review, 14*(2), 187-193. doi:http://dx.doi.org/10.3758/BF03194050

Paul, A. M. (2015). Researchers find that frequent tests can boost learning. *Scientific American.* Available at: http://www.scientificamerican.com/article/researchers-find-that-frequent-tests-can-boost-learning/

Perkins, D. N., & Salomon, G. (1994). Transfer of learning. In T. Husen & T. N. Postelwhite (Eds.). *International Handbook of Educational Research* (Second Edition, Vol. 11; pp. 6452-6457). Oxford: Pergamon Press.

*Peterson, D. J., & Mulligan, N. W. (2013). The negative testing effect and multifactor account. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 39*(4), 1287-1293. doi:http://dx.doi.org/10.1037/a0031337

*Pilotti, M., Chodorow, M., & Petrov, R. (2009). The usefulness of retrieval practice and review-only practice for answering conceptually related test questions. *Journal of General Psychology, 136*(2), 179-203. doi:http://dx.doi.org/10.3200/GENP.136.2.179-204

Popham, W. J. (2011). *Classroom assessment: What teachers need to know* (6th edn.). Boston,: Allyn & Bacon.

Pyc, M. A., & Rawson, K. A. (2010). Why testing improves memory: Mediator effectiveness hypothesis. *Science, 330*(6002), 335.

Raaijmakers, J. G., & Shiffrin, R. M. (1981). Search of associative memory. *Psychological Review, 88*(2), 93-134. doi:http://dx.doi.org/10.1037/0033-295X.88.2.93

Raudenbush, S. W. (2009). Analyzing effect sizes: Random-effects models. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *Handbook of research synthesis and meta-analysis* (2nd

ed., pp. 295–315). New York: Russell Sage Foundation.

Rawson, K. A., & Dunlosky, J. (2011). Optimizing schedules of retrieval practice for durable and efficient learning: How much is enough? *Journal of Experimental Psychology: General, 140*(3), 283-302. doi:http://dx.doi.org/10.1037/a0023956

Rawson, K. A., & Dunlosky, J. (2012). When is practice testing most effective for improving the durability and efficiency of student learning? *Educational Psychology Review, 24*(3), 419-435. doi:http://dx.doi.org/10.1007/s10648-012-9203-1

Rawson, K. A., Dunlosky, J., & Sciartelli, S. M. (2013). The power of successive relearning: Improving performance on course exams and long-term retention. *Educational Psychology Review, 25*(4), 523-548. doi:http://dx.doi.org/10.1007/s10648-013-9240-4

*Rawson, K. A., Vaughn, K. E., & Carpenter, S. K. (2015). Does the benefit of testing depend on lag, and if so, why? evaluating the elaborative retrieval hypothesis. *Memory & Cognition, 43*(4), 619-633. doi:http://dx.doi.org/10.3758/s13421-014-0477-z

Rickard, T. C., & Bourne, L. E., Jr. (1996). Some tests of an identical elements model of basic arithmetic skills. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22*(5), 1281-1295. doi:http://dx.doi.org/10.1037/0278-7393.22.5.1281

Rickard, T. C., Healy, A. F., & Bourne, L. E. (1994). On the cognitive structure of basic arithmetic skills: Operation, order, and symbol transfer effects. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20*(5), 1139-1153. doi:http://dx.doi.org/10.1037/0278-7393.20.5.1139

Rickard, T. C., & Pan, S. C. (2017). A dual memory theory of the retrieval practice effect. *Psychonomic Bulletin & Review.* doi:http://dx.doi.org/10.3758/s13423-017-1298-4

*Rickard, T. C., & Pan, S. C. *Test-enhanced learning of pairs, triplets, and facts: when and why*

does transfer occur? Unpublished manuscript.

Rickard, T. C., Healy, A. F., & Bourne, L. E. (1994). On the cognitive structure of basic arithmetic skills: Operation, order, and symbol transfer effects. *Journal of Experimental Psychology: Learning, Memory, and Cognition,* 20(5), 1139-1153. doi:http://dx.doi.org/10.1037/0278-7393.20.5.1139

Roediger, H. L. (2007). Transfer: the ubiquitous concept. In H. L. Roediger, Y. Dudai & S. M. Fitzpatrick (Eds.), *Science of memory: Concepts,* New York: Oxford University Press.

Roediger, H. L., & Blaxton, T. A. (1987). Retrieval modes produce dissociations in memory for surface information. *Memory and Learning: The Ebbinghaus Centennial Conference*. (pp. 349-379). Hillsdale: Lawrence Erlbaum Associates, Inc.

Roediger, H. L., & Butler, A. C. (2011). The critical role of retrieval practice in long-term retention. *Trends in Cognitive Sciences*, 15(1), 20-27. doi:http://dx.doi.org/10.1016/j.tics.2010.09.003

Roediger, H. L., & Karpicke, J. D. (2006). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science, 1*(3), 181-210. doi:http://dx.doi.org/10.1111/j.1745-6916.2006.00012.x

Roediger, H. L., Putnam, A. L., & Smith, M. A. (2011). Ten benefits of testing and their applications to educational practice. *The psychology of learning and motivation (vol 55): Cognition in education*. (pp. 1-36). San Diego: Elsevier Academic Press. doi:http://dx.doi.org/10.1016/B978-0-12-387691-1.00001-6

Roediger, H. L., & Pyc, M. A. (2012). Inexpensive techniques to improve education: Applying cognitive psychology to enhance educational practice. *Journal of Applied Research in Memory and Cognition, 1*(4), 242-248. doi:http://dx.doi.org/10.1016/j.jarmac.2012.09.002

Roediger, H. L., & Marsh, E. J. (2005). The positive and negative consequences of multiple-choice testing. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31*(5), 1155-1159. doi:http://dx.doi.org/10.1037/0278-7393.31.5.1155

*Rohrer, D., Taylor, K., & Sholar, B. (2010). Tests enhance the transfer of learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition,* 36(1), 233-239. doi:http://dx.doi.org/10.1037/a0017678

Rowland, C. A. (2014). The effect of testing versus restudy on retention: A meta-analytic review of the testing effect. *Psychological Bulletin, 140*(6), 1432-1463. doi:http://dx.doi.org/10.1037/a0037559

*Rowland, C. A., & DeLosh, E. L. (2015). Mnemonic benefits of retrieval practice at short retention intervals. *Memory, 23*(3), 403-419. doi:http://dx.doi.org/10.1080/09658211.2014.889710

*Rowland, C. A., Littrell-Baez, M., Sensenig, A. E., & DeLosh, E. L. (2014). Testing effects in mixed- versus pure-list designs. *Memory & Cognition,* 42(6), 912-921. doi:http://dx.doi.org/10.3758/s13421-014-0404-3

Runquist, W. N. (1983). Some effects of remembering on forgetting. *Memory & Cognition, 11*(6), 641-650.

Salomon, G., & Perkins, D. N. (1989). Rocky roads to transfer: Rethinking mechanisms of a neglected phenomenon. *Educational Psychologist,* 24(2), 113-142.

Schmidt, R. A., & Bjork, R. A. (1992). New conceptualizations of practice: Common principles in three paradigms suggest new concepts for training. *Psychological Science, 3*(4), 207-217. doi:http://dx.doi.org/10.1111/j.1467-9280.1992.tb00029.x

*Sensenig, A. E. (2011). *Multiple choice testing and the retrieval hypothesis of the testing effect* (Order No. AAI3428630).

*Sensenig, A. E., Littrell-Baez, M., & DeLosh, E. L. (2011). Testing effects for common versus proper names. *Memory,* 19(6), 664-673. doi:http://dx.doi.org/10.1080/09658211.2011.599935

Singley, M. K., & Anderson, J. R. (1989). *The transfer of cognitive skill.* Cambridge: Harvard University Press.

Smith, M. A., & Karpicke, J. D. (2014). Retrieval practice with short-answer, multiple-choice, and hybrid tests. *Memory,* 22(7), 784-802. doi:http://dx.doi.org/10.1080/09658211.2013.831454

Smithson, M. (2003). *Confidence Intervals*. SAGE.

Soderstrom, N. C., & Bjork, R. A. (2014). Testing facilitates the regulation of subsequent study time. *Journal of Memory and Language, 73*, 99-115. doi:http://dx.doi.org/10.1016/j.jml.2014.03.003

Stanislaw, H., & Todorov, N. (1999). Calculation of signal detection theory measures. *Behavior Research Methods, Instruments & Computers, 31*(1), 137-149. doi:http://dx.doi.org/10.3758/BF03207704

Stanley, T. D. (2017). Limitations of PET-PEESE and other meta-analysis methods. *Social Psychological and Personality Science*, 1948550617693062.

Stanley, T. D., & Doucouliagos, H. (2014). Meta-regression approximations to reduce publication selection bias. *Research Synthesis Methods, 5*(1), 60-78. doi:http://dx.doi.org/10.1002/jrsm.1095

Sternberg, R. (1993). *Transfer on trial: Intelligence, cognition, and instruction*. Westport: Ablex Publishing.

Strube, M. J., & Hartmann, D. P. (1983). Meta-analysis: Techniques, applications, and functions. *Journal of Consulting and Clinical Psychology, 51*(1), 14-27.

Swaminathan, N. (2006). Testing improves retention—even of material not on exam. *Scientific American*. Available at: http://www.scientificamerican.com/article/testing-improves-retentio/

Tanner-Smith, E. E., & Tipton, E. (2014). Robust variance estimation with dependent effect sizes: Practical considerations including a software tutorial in Stata and SPSS. *Research Synthesis Methods*, 5(1), 13-30. doi: 10.1002/jrsm.1091

Thorndike, E. L. (1906). *The elements of psychology.* New York: A.G. Seiler.

*Tran, R., Rohrer, D., & Pashler, H. (2015). Retrieval practice: The lack of transfer to deductive inferences. *Psychonomic Bulletin & Review,* 22(1), 135-140. doi:http://dx.doi.org/10.3758/s13423-014-0646-x

Tullis, J. G., Benjamin, A. S., & Ross, B. H. (2014). The reminding effect: Presentation of associates enhances memory for related words in a list. *Journal of Experimental Psychology: General, 143*(4), 1526-1540. doi:http://dx.doi.org/10.1037/a0036036

Tulving, E. (1984). Précis of elements of episodic memory. *Behavioral and Brain Sciences, 7*(2), 223-268.

Van den Broek, G., Takashima, A., Wiklund-Hörnqvist, C., Wirebring, L. K., Segers, E., Verhoeven, L., & Nyberg, L. (2016). Neurocognitive mechanisms of the "testing effect": a review. *Trends in Neuroscience and Education*, *5*(2), 52-66.

Van den Bussche, E., Van den Noortgate, W., & Reynvoet, B. (2009). Mechanisms of masked priming: A meta-analysis. *Psychological Bulletin, 135*(3), 452-477. doi:http://dx.doi.org/10.1037/a0015329; 10.1037/a0015329

*van Eersel, G. G., Verkoeijen, P. P., Povilenaite, M., & Rikers, R. (2016). The testing effect and far transfer: the role of exposure to key information. *Frontiers in Psychology*, *7*

*van Gog, T., & Kester, L. (2012). A test of the testing effect: Acquiring problem-solving skills

from worked examples. *Cognitive Science,* 36(8), 1532-1541. doi:http://dx.doi.org/10.1111/cogs.12002

van Gog, T., & Paas, F. (2008). Instructional efficiency: Revisiting the original construct in educational research. *Educational Psychologist, 43*(1), 16-26. doi:http://dx.doi.org/10.1080/00461520701756248

van Gog, T., & Sweller, J. (2015). Not new, but nearly forgotten: The testing effect decreases or even disappears as the complexity of learning materials increases. *Educational Psychology Review,* 27(2), 247-264. doi:http://dx.doi.org/10.1016/j.cedpsych.2010.10.004

*van Gog, T., Kester, L., & Paas, F. (2011). Effects of worked examples, example-problem, and problem-example pairs on novices' learning. *Contemporary Educational Psychology,* 36(3), 212-218. doi:http://dx.doi.org/10.1016/j.cedpsych.2010.10.004

*van Gog, T., Kester, L., Dirkx, K., Hoogerheide, V., Boerboom, J., & Verkoeijen, P. P. J. L. (2015). Testing after worked example study does not enhance delayed problem-solving performance compared to restudy. *Educational Psychology Review*, 27(2), 265-289 doi:http://dx.doi.org/10.1007/s10648-015-9297-3

Vaughn, K. E., & Rawson, K. A. (2014). Effects of criterion level on associative memory: Evidence for associative asymmetry. *Journal of Memory and Language, 75*, 14-26. doi:http://dx.doi.org/10.1016/j.jml.2014.04.004

*Veltre, M. T., Cho, K. W., & Neely, J. H. (2015). Transfer-appropriate processing in the testing effect. *Memory, 23*(8), 1229-1237. doi:http://dx.doi.org/10.1080/09658211.2014.970196

*Verkoeijen, P. P. J. L., Bouwmeester, S., & Camp, G. (2012). A short-term testing effect in cross-language recognition. *Psychological Science,* 23(6), 567-571. doi:http://dx.doi.org/10.1177/0956797611435132

Vevea, J. L., & Hedges, L. V. (1995). A general linear model for estimating effect size in the presence of publication bias. *Psychometrika, 60*(3), 419-435.

Vevea, J. L., & Woods, C. M. (2005). Publication bias in research synthesis: sensitivity analysis using a priori weight functions. *Psychological Methods*, *10*(4), 428.

Wenger, S. K., Thompson, C. P., & Bartling, C. A. (1980). Recall facilitates subsequent recognition. *Journal of Experimental Psychology: Human Learning and Memory,* 6(2), 135-144. doi:http://dx.doi.org/10.1037/0278-7393.6.2.135

Wissman, K. T., Rawson, K. A., & Pyc, M. A. (2011). The interim test effect: Testing prior material can facilitate the learning of new material. *Psychonomic Bulletin & Review, 18*(6), 1140-1147. doi:http://dx.doi.org/10.3758/s13423-011-0140-7

*Wooldridge, C. L., Bugg, J. M., McDaniel, M. A., & Liu, Y. (2014). The testing effect with authentic educational materials: A cautionary note. *Journal of Applied Research in Memory and Cognition,* 3(3), 214-221. doi:http://dx.doi.org/10.1016/j.jarmac.2014.07.001

Wuensch, K. L. (2012). *Using SPSS to obtain a confidence interval for Cohen's d*. Available at: http://core.ecu.edu/psyc/wuenschk/SPSS/CI-d-SPSS.pdf.

Wylie, H. H. (1919). An experimental study of transfer of response in the white rat. *Behavior Monographs, 3.*

Yeo., D., & Fazio, L. (in press). *The optimal learning strategy depends on learning goals and processes: retrieval practice vs. worked examples*. *Journal of Educational Psychology.*

*Zhou, A., Ma, X., Li, J., & Cui, D. (2013). The advantage effect of retrieval practice on memory retention and transfer: Based on explanation of cognitive load theory. *Acta Psychologica Sinica, 45*(8), 849-859. doi:http://dx.doi.org/10.3724/SP.J.1041.2013.00849

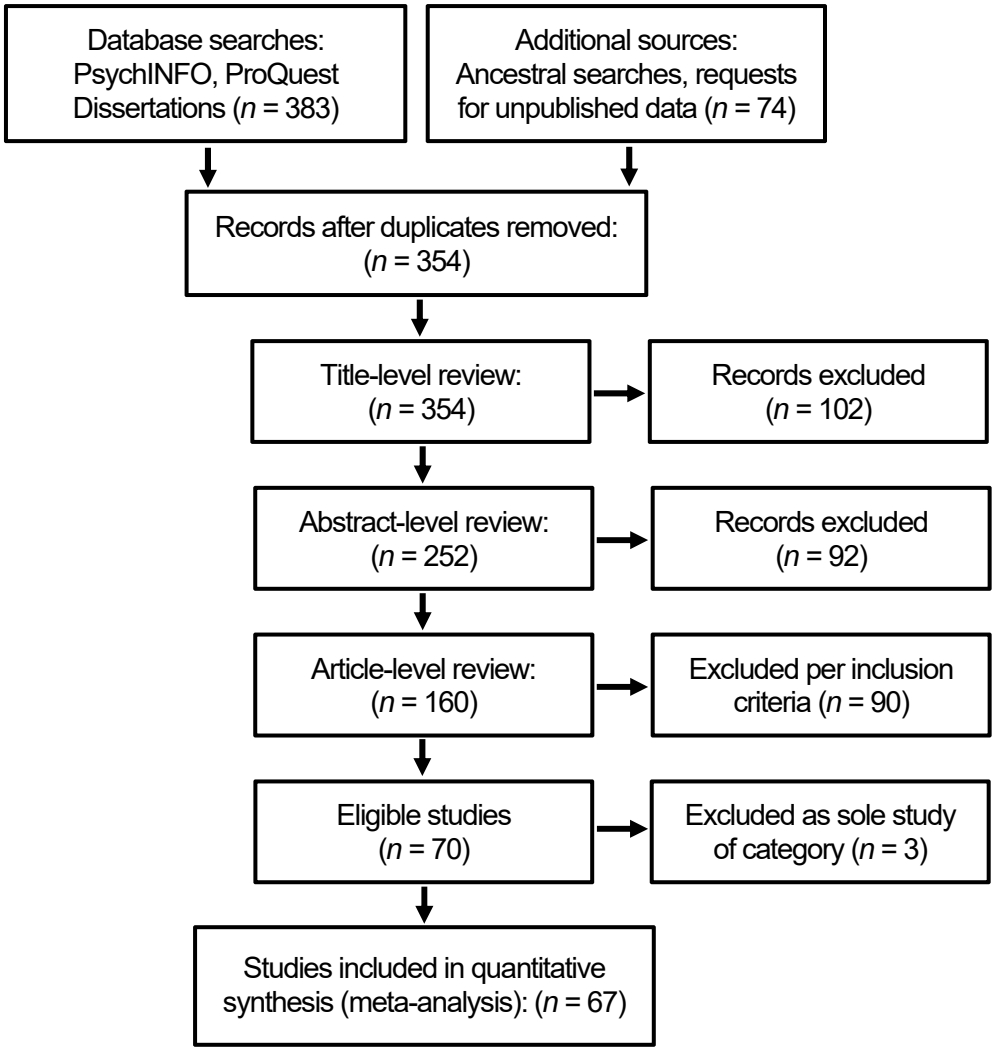

Figure 1. Flowchart of the literature search and selection process (*n* refers to individual studies).

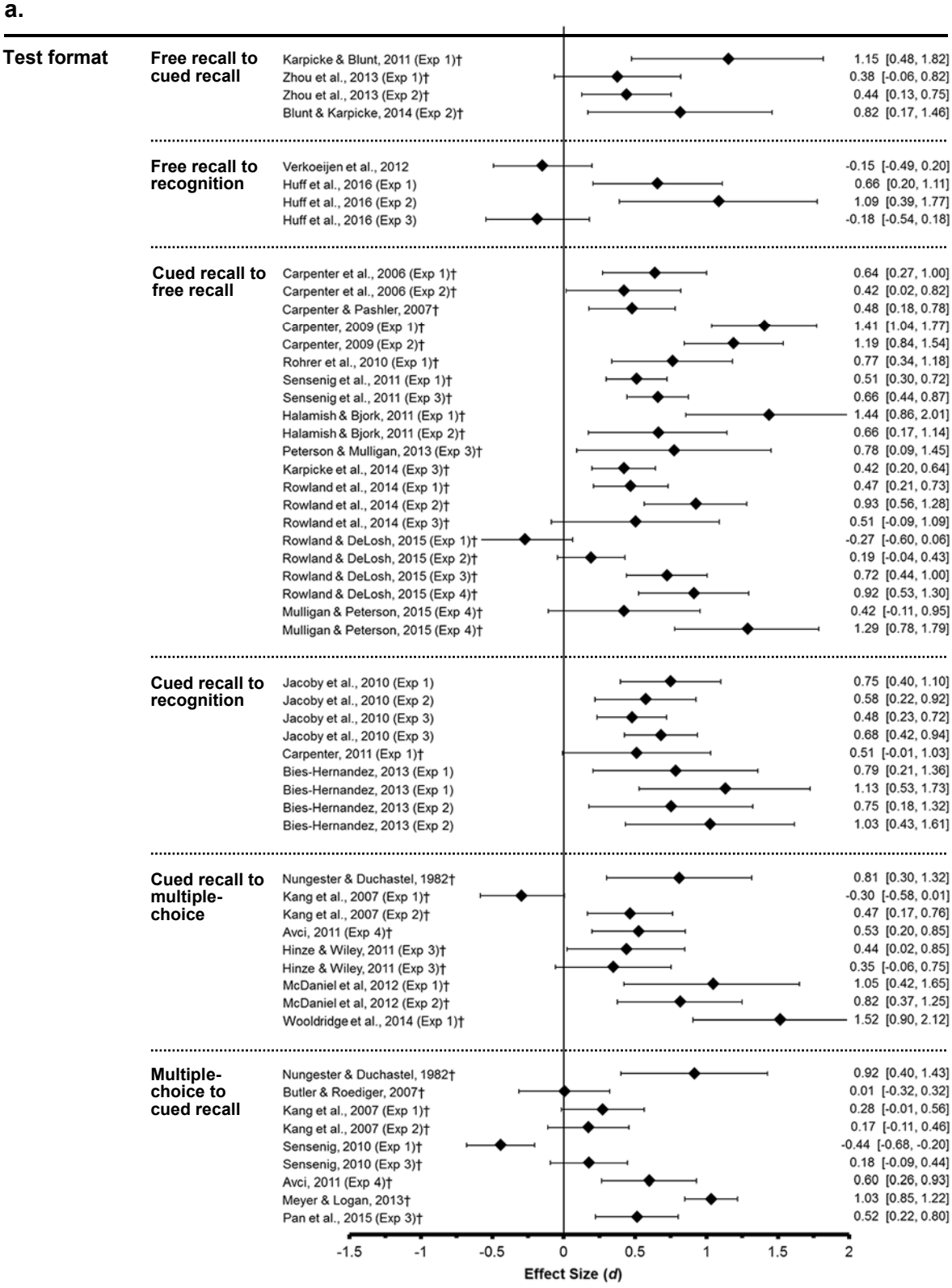

Figure 2, panel a. Forest plot of effect sizes (Cohen's *d*) with 95% confidence intervals for the transfer across test formats category. Study order matches Table 1. (*) denotes exclusions due to non-independent reexposure controls; (†) denotes strong response congruency and/or elaborated retrieval practice.

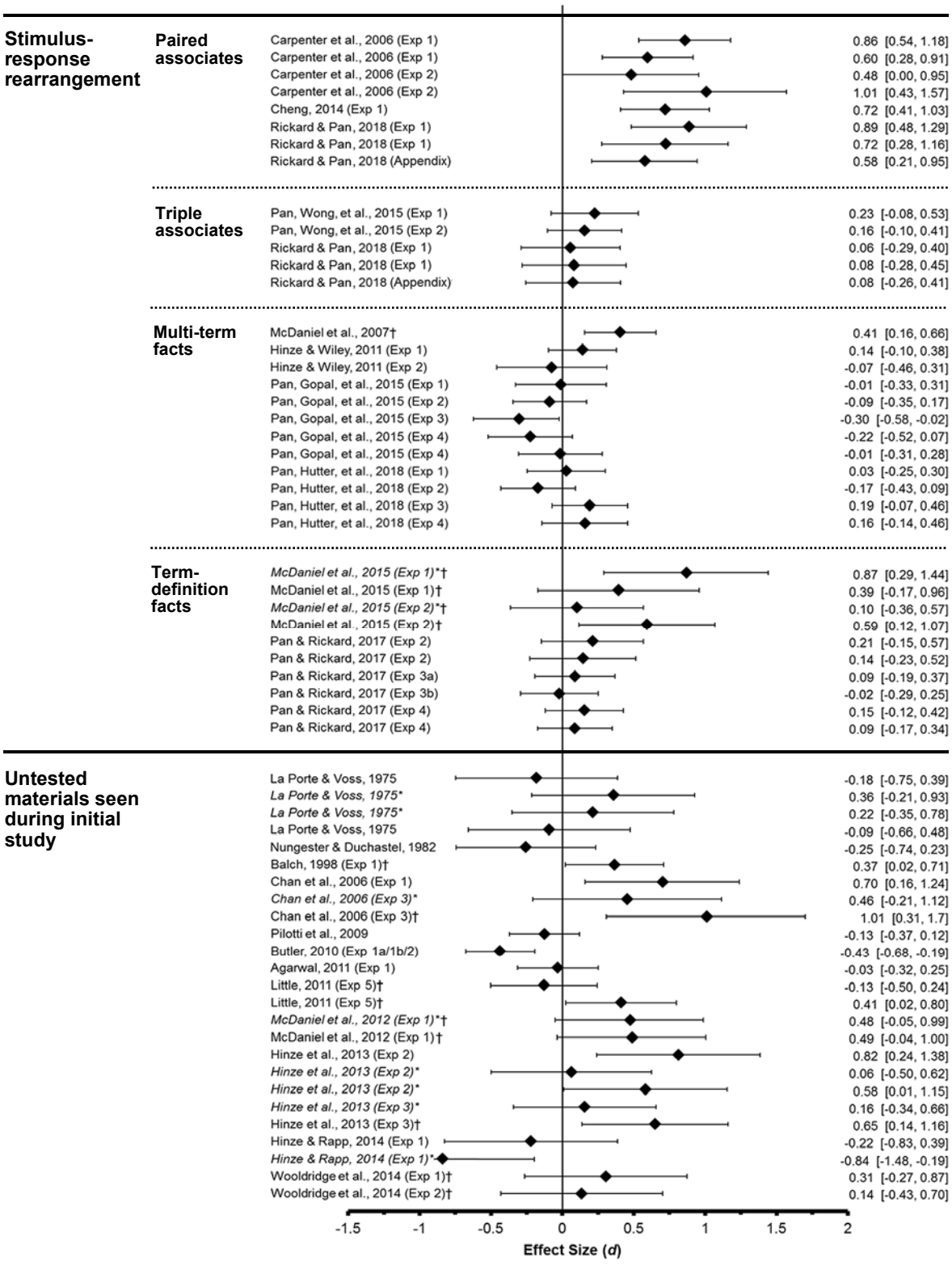

Figure 2, panel b. Forest plot of effect sizes (Cohen's *d*) with 95% confidence intervals for the transfer to stimulus-response rearrangement and to untested materials categories. Study order matches Table 1. (*) denotes exclusions due to non-independent reexposure controls; (†) denotes strong response congruency and/or elaborated retrieval practice.

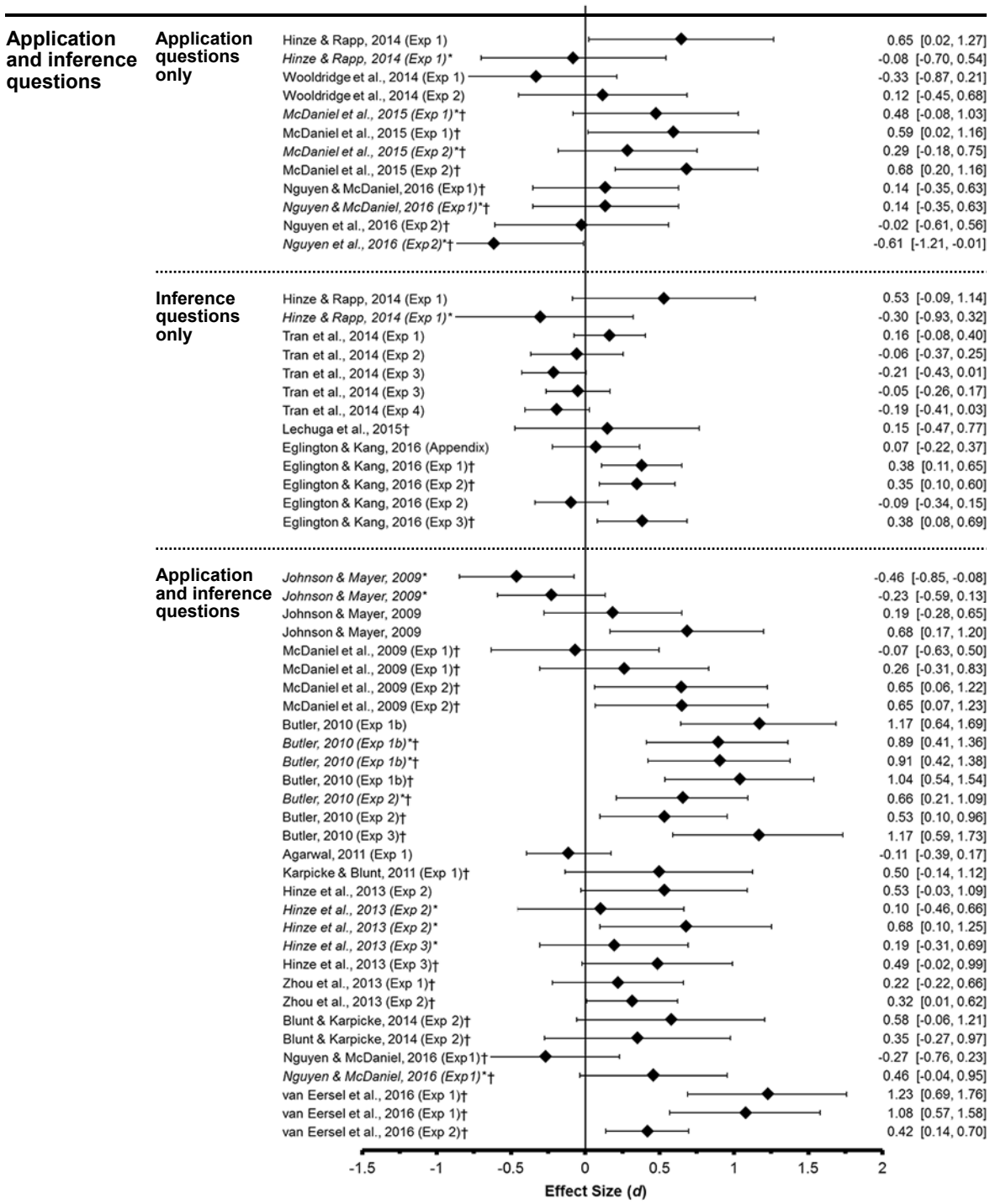

Figure 2, panel c. Forest plot of effect sizes (Cohen's *d*) with 95% confidence intervals for the transfer to application and inference questions category. Study order corresponds to Table 1. (*) indicates exclusions due to non-independent reexposure controls; (†) indicates strong response congruency and/or elaborated retrieval practice.

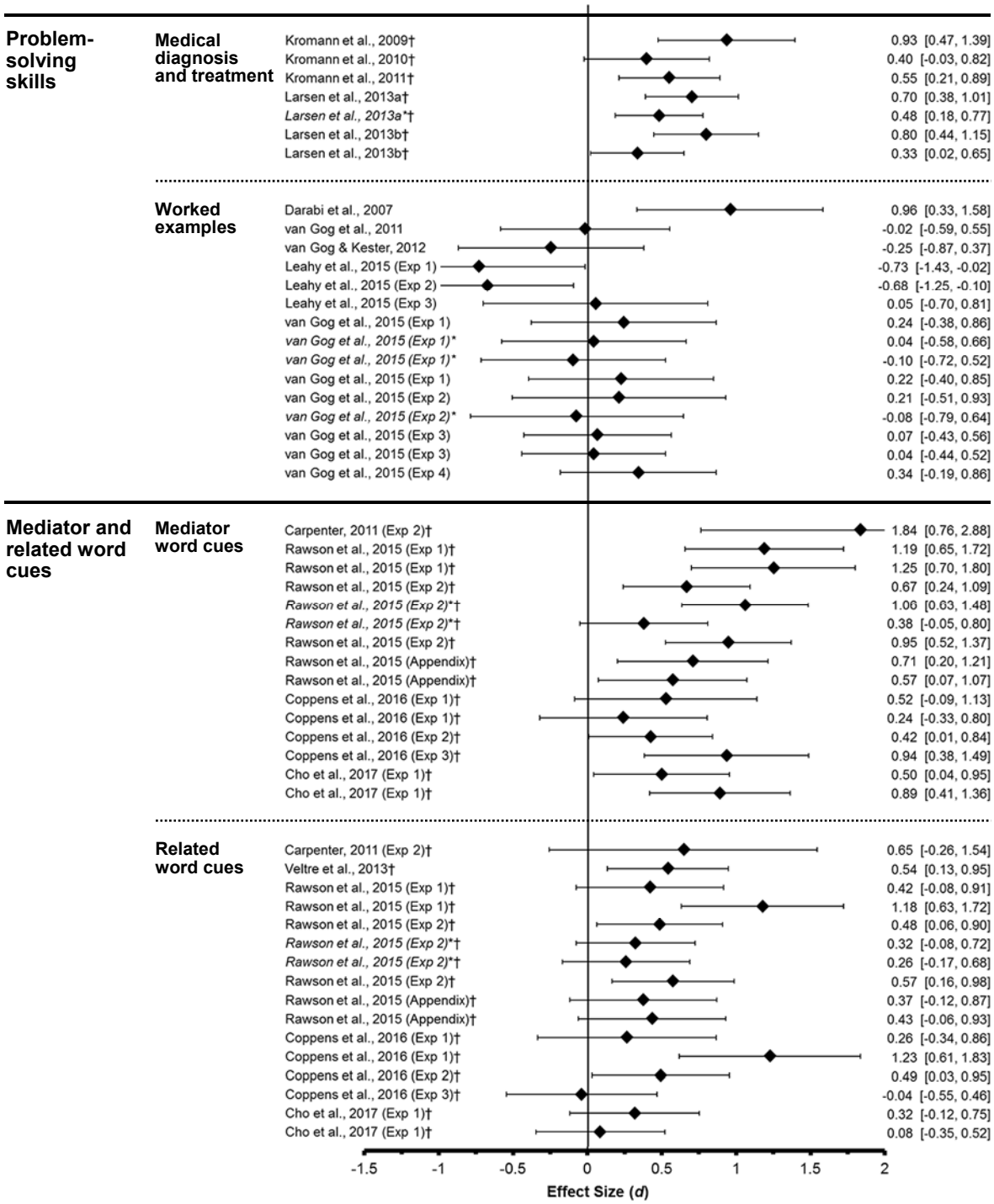

Figure 2, panel d. Forest plot of effect sizes (Cohen's *d*) with 95% confidence intervals for the transfer of problem-solving skills and mediator and related word cues categories. Study order corresponds to Table 1. (*) indicates exclusions due to non-independent reexposure controls; (†) indicates strong response congruency and/or elaborated retrieval practice.

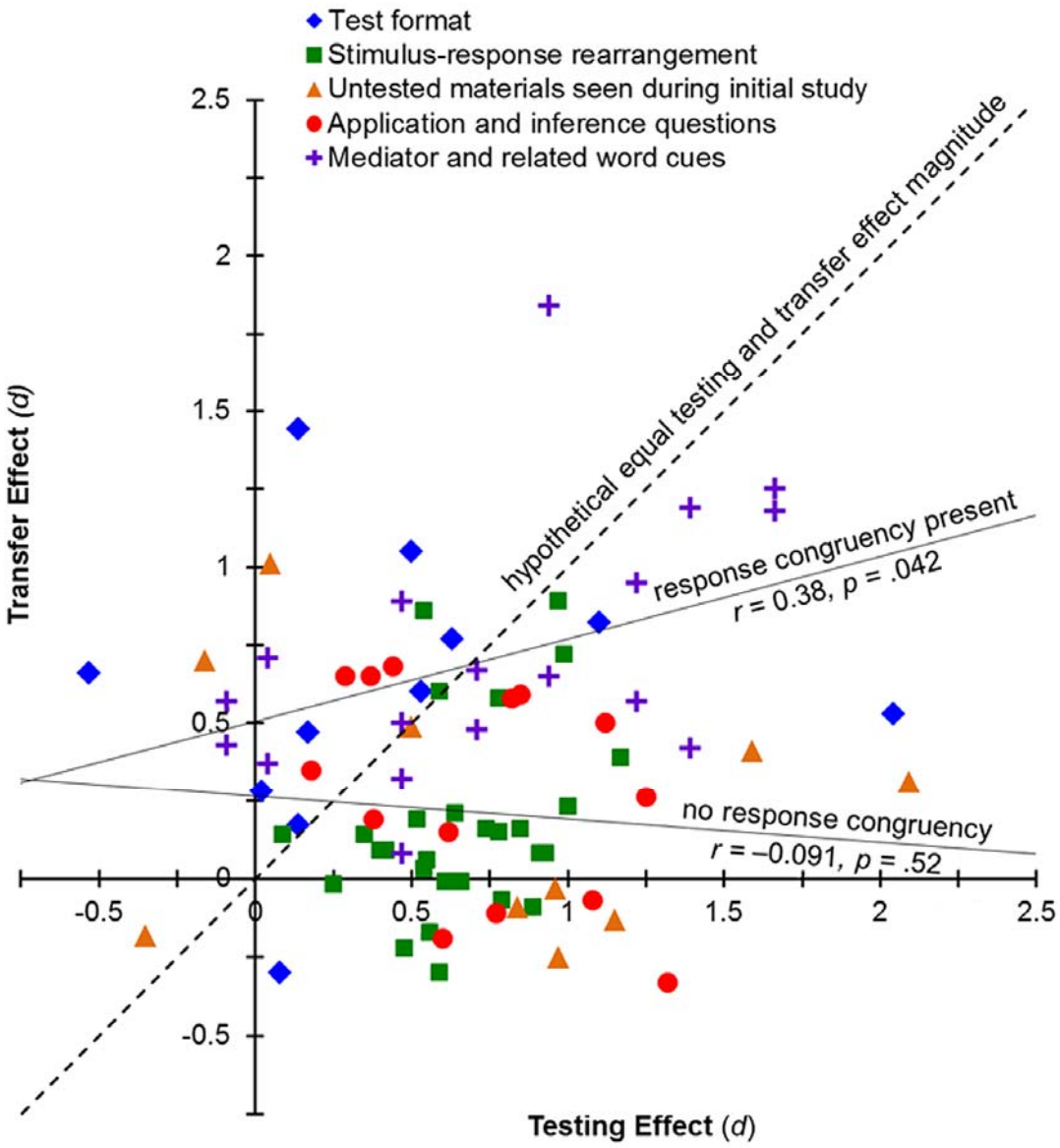

Figure 3. Scatterplot of testing vs. transfer effect sizes (Cohen's *d*), from 28 studies that assessed both effects from within the same experiments (there were 81 such cases; all categories except problem-solving skills are represented). The dotted line represents the hypothetical case of equal testing and transfer effect magnitude (for points above the line, the transfer effect is larger; for the reverse case, the testing effect is larger). The solid diagonal lines represent the best least squares regression fit to data for categories with strong response congruency (i.e., test format and mediator and related cues) and no response congruency (all other categories).

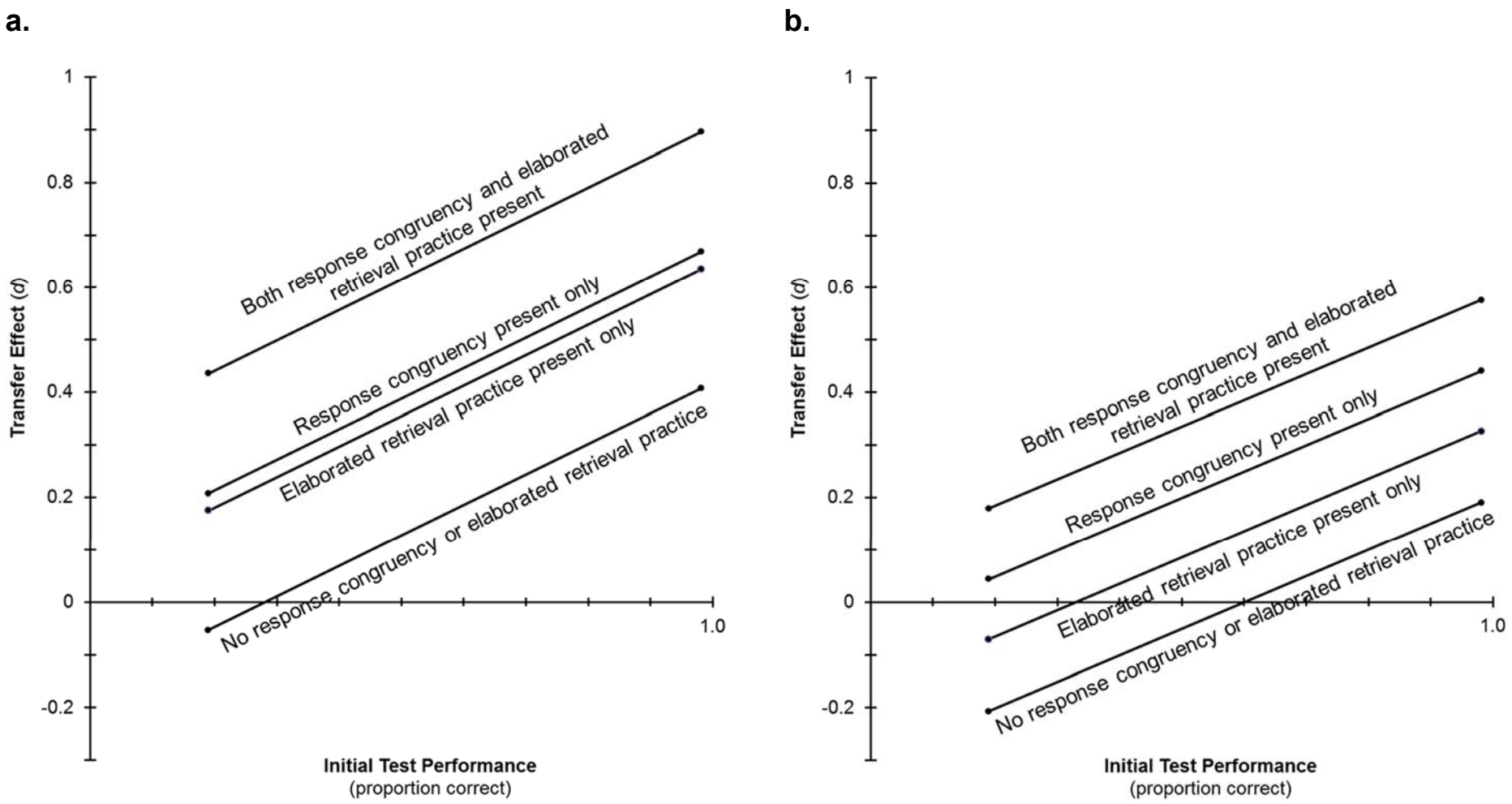

Figure 4. A *three-factor framework* for transfer of test-enhanced learning. Lines represent predicted transfer effect sizes as a function of three factors: initial test performance and the presence or absence of response congruency and elaborated retrieval practice. Panel a: random-effects analysis estimates. Panel b: PEESE analysis estimates. Effect size estimates are drawn across the full range of proportion correct in the dataset (which was from 0.19 to 0.98).

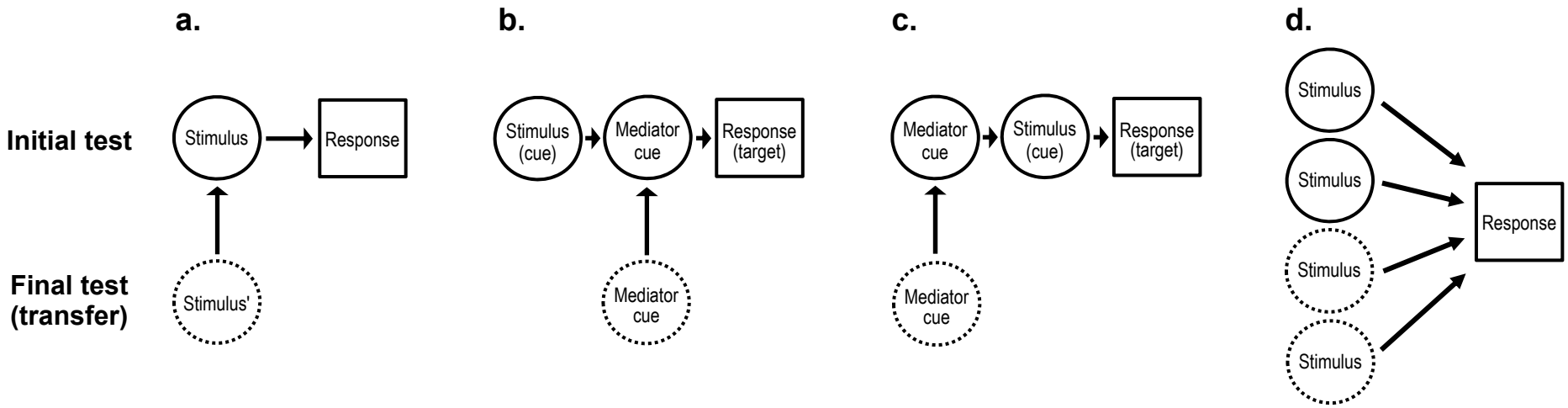

Figure 5. Four scenarios of positive transfer via response congruency and stimulus-to-response pathway reinstatement. Solid arrows represent proposed pathways. Panel a: the transfer stimulus, *stimulus'*, is a minimally modified version of the stimulus on the initial test (as occurs in many cases of transfer across test formats, e.g., multiple-choice to cued recall as in Pan, Gopal, et al., 2015). Panel b: partial pathway reinstatement for mediator cues on the transfer test, wherein an associative pathway that was directly formed between the mediator and the target on the initial test is reactivated (Coppens et al., 2016). Panel c: full stimulus-to-response pathway reinstatement for mediator cues on the transfer test wherein the mediator prompts the original cue and leads to recall of the target (Cho et al., 2017). Panel d: the stimuli on the transfer test are a subset of, or overlap with, the stimuli on the initial test (as may occur for some cases of problem-solving involving medical diagnosis and treatment, e.g., Larsen et al., 2016a).

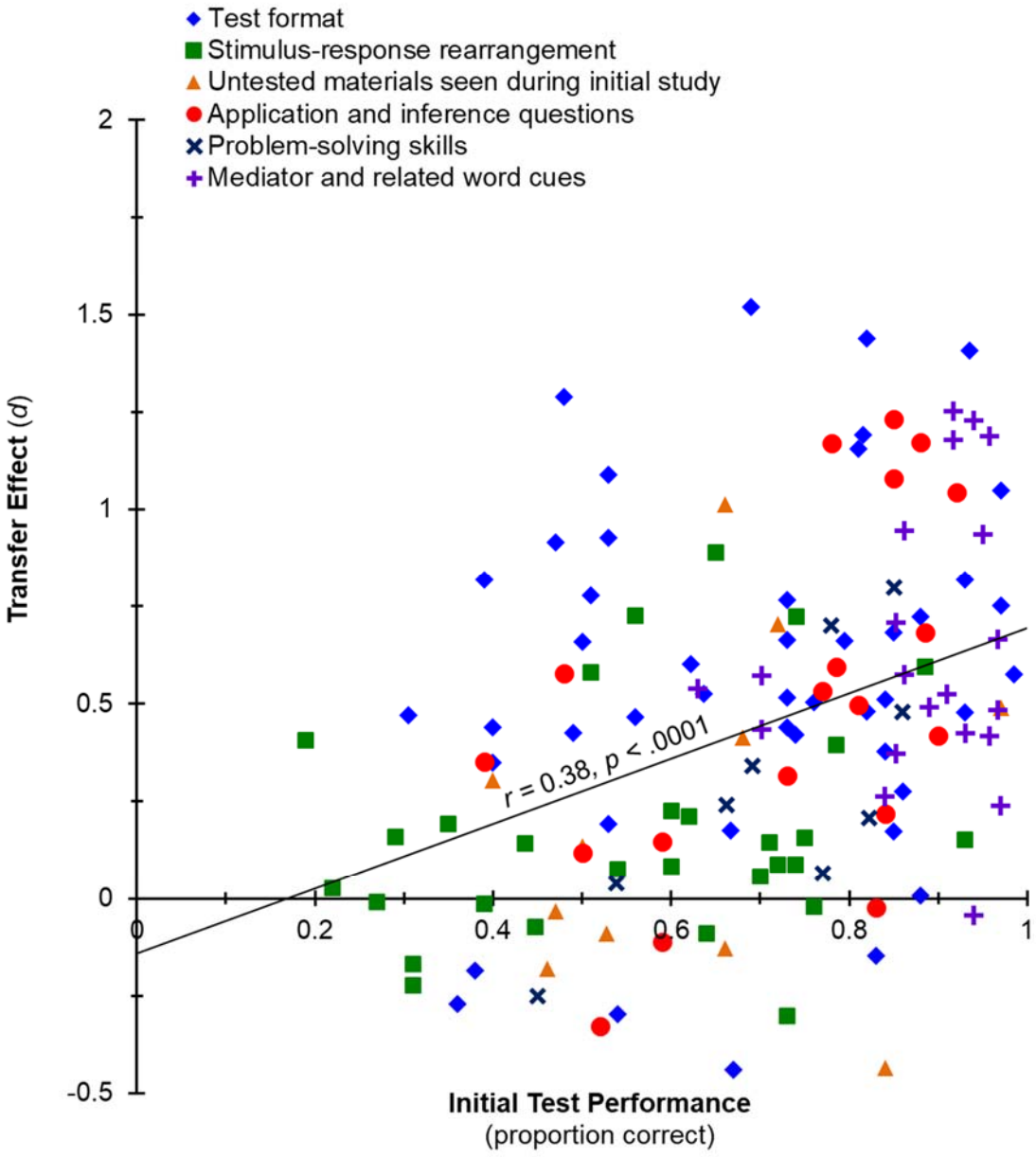

Figure 6. Scatterplot depicting the relationship between transfer effect size (Cohen's *d*) and initial test performance (proportion correct), where reported ($k = 135$ effect sizes), across all six major transfer categories. The diagonal line represents the best least squares regression fit to the data.

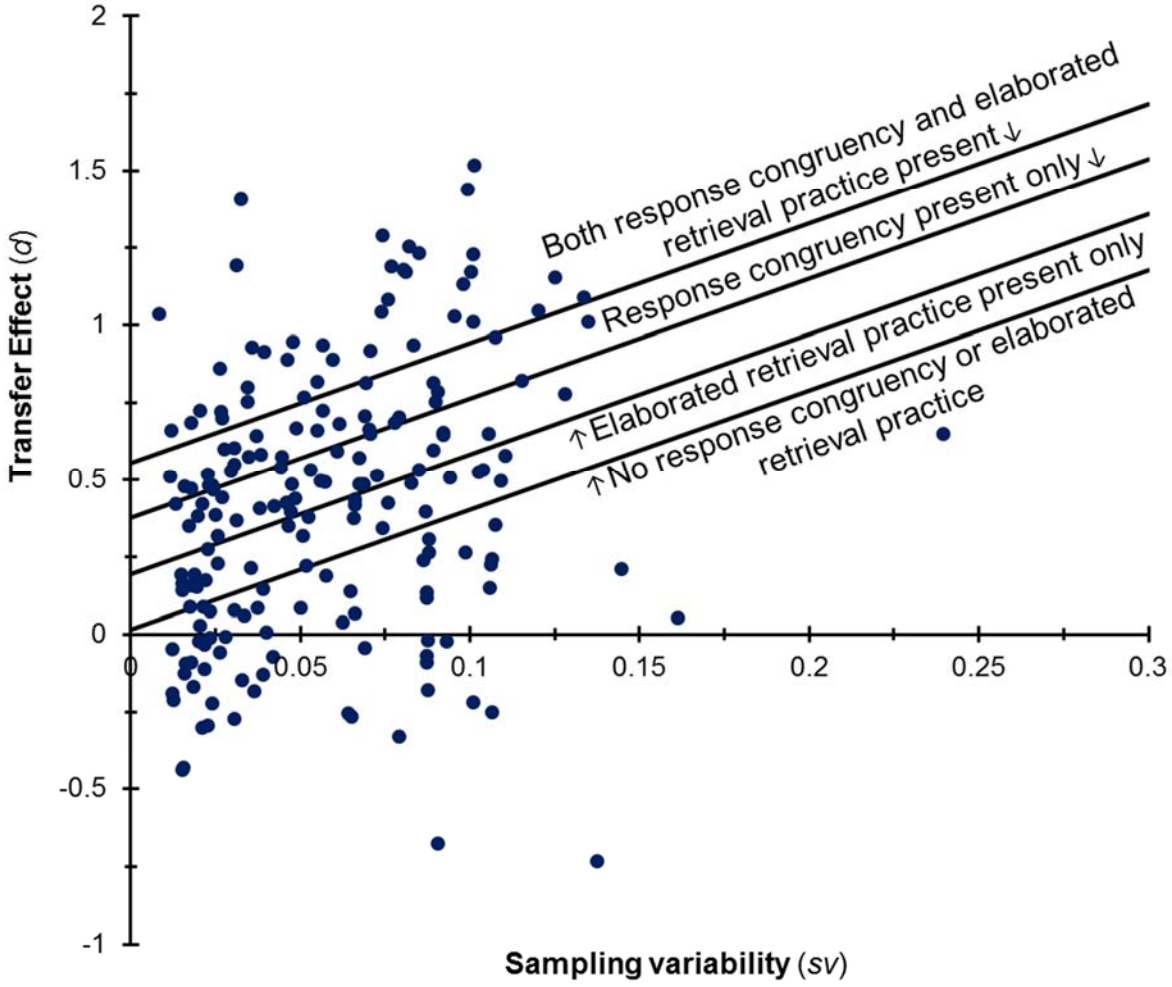

Figure 7. All 192 transfer effect sizes (Cohen's *d*) in the overall dataset plotted as a function of sampling variability (*sv*). PEESE analysis estimates for the different levels and combinations of response congruency and elaborated retrieval practice are also plotted. The intercepts of the plotted moderator lines depict the estimated effect size for each moderator when *sv* is at a hypothetical value of zero, and hence in principle a state of no publication bias. Inspection of the scatterplot suggests publication bias; as *sv* increases, the upper half of the plot has more effect sizes than the lower half.

Table 1.

Studies of Transfer of Test-Enhanced Learning

				Study design											
Category	Sub-category	Reference	Exp.	Stimuli	Feedback	Delay (hr)	Train	Final	Initial test condition	Reexposure control	Transfer test	N_T	N_R	*d*	*sv*
Test format															
	Free recall to cued recall														
		Karpicke & Blunt, 2011	1	Passage	Y (R)	168	FR	CR	RP	Rpt. study	Verbatim	20	20	1.15	0.13
		Zhou et al., 2013	1	Passage	Y (R)	168	FR	CR	RP	Restudy	Factual	40	40	0.38	0.052
			2	Passage	Y (R)	168	FR	CR	RP	Concept mapping	Factual	43		0.44	0.027
		Blunt & Karpicke, 2014	2	Passage	Y (R)	168	FR	CR	RP, paragraph	Rpt. study, paragraph	Verbatim	20	20	0.82	0.12
	Free recall to recognition														
		Verkoeijen et al., 2012		Words	N	0.033	FR	Recog	Testing	Restudy	Within language	33		-0.15	0.033
		Huff et al., 2016	1	Words	N	0	FR	Recog	Recall, list	Restudy	Corrected recog.	39	40	0.66	0.055
			2	Words	N	33	FR	Recog	Recall, list	Restudy	Corrected recog.	18	19	1.09	0.13
			3	Words	N	33	FR	Recog	Recall, list	Restudy	Corrected recog.	30		-0.18	0.036
	Cued recall to free recall														
		Carpenter et al., 2006	1	PAL	Y	33	CR	FR	Test/study	Study	Recall bs	35		0.64	0.037
			2	PAL	Y	33	CR	FR	Test/study	Study	Recall bs	26		0.42	0.046
		Carpenter & Pashler, 2007		Maps	Y	0.5	CR	FR	Test/study	Study	LQ, abs. acc.	50		0.48	0.023
		Carpenter, 2009	1	PAL	N	0.083	CR	FR	Test	Study		60		1.41	0.032
			2	PAL	N	0.083	CR	FR	Test	Study	Final test	76	76	1.19	0.031
		Rohrer et al., 2010	1	Maps	Y	24	CR	FR	TS	Study-only	Transfer test	28		0.77	0.051
		Sensenig et al., 2011	1	Words	N	0.083	CR	FR	Tested, occupation/name	Restudied	Final test	98		0.51	0.012
			3	Words	N	0.083	CR	FR	Tested, name/noun	Restudied	Final test	103		0.66	0.012
		Halamish & Bjork, 2011	1	PAL	N	0.14	CR	FR	STT	SSS	Difficult test	24		1.44	0.10
			2	PAL	N	0.14	CR	FR	STT	SSS	Difficult test	20		0.66	0.070
		Peterson & Mulligan, 2013	3	PAL	Y	0	CR	FR	Retrieval	Restudy	Final test	18	18	0.78	0.13
		Karpicke et al., 2014	3	Texts	N	0	CR	FR	Guided retrieval	Reread	Final test	85		0.42	0.013
		Rowland et al., 2014	1	Words	N	0.083	CR	FR	Test, pure+mixed	Restudy, pure+mixed	Final test	64		0.47	0.018
			2	Words	N	0.083	CR	FR	Test, pure+mixed	Restudy, pure+mixed	Final test	43		0.93	0.036
			3	Words	N	0.067	CR	FR	Test, pure list	Restudy, pure list	Final test	23	23	0.51	0.094
		Rowland & DeLosh, 2015	1	Words	N	0.075	CR	FR	Tested	Restudy	Final test	36		-0.27	0.031
			2	Words	N	0.075	CR	FR	Tested	Restudy	Final test	71		0.19	0.015
			3	Words	N	0.038	CR	FR	Tested	Restudy	Final test	63		0.72	0.020
			4	Words	Y	0.017	CR	FR	Tested	Restudy	Final test	38		0.92	0.039
		Mulligan & Peterson, 2015	4	PAL	Y	0.25	CR	FR	Retrieval, pure list	Restudy	Final test	28	28	0.42	0.076
			4	PAL	Y	0.25	CR	FR	Retrieval, mixed list	Restudy	Final test	28		1.29	0.074

(table continues)

Table 1 (*continued*)

				Study design											
Category	Sub-category	Reference	Exp.	Stimuli	Feedback	Delay (hr)	Train	Final	Initial test condition	Reexposure control	Transfer test	N_T	N_R	*d*	*sv*
	Cued recall to recognition														
		Jacoby et al., 2010	1	CAT	Y	0	CR	Recog.	$ST_sT_sT_s$	SSSS	Hit rate	40		0.75	0.034
			2	CAT	Y	0	CR	Recog.	$ST_sT_sT_sT_sT_s$	SSSSSS	Hit rate	36		0.58	0.035
			3	CAT	N	12	CR	Recog	SSSTTT‡	SSSSSS	Hit rate	72		0.48	0.016
			3	CAT	N	12	CR	Recog	STTTTT‡	SSSSSS	Hit rate	72		0.68	0.018
		Carpenter, 2011	1	PAL	N	0.083	CR	Recog	Test	Study	Hit rate targets	30	29	0.51	0.073
		Bies-Hernandez, 2013	1	Words	Y	48	CR	Recog	ST	Restudying	Final test	25	25	0.79	0.090
			1	Words	Y	48	CR	Recog	STTT	Restudying	Final test	25	25	1.13	0.098
			2	Words	Y	48	CR	Recog	ST	Restudying	Standard	25	25	0.75	0.090
			2	Words	Y	48	CR	Recog	ST	Restudying	Recollection	25	25	1.03	0.095
	Cued recall to multiple-choice														
		Nungester & Duchastel, 1982		Passages	N	336	CR	MC	Test	Review	Old items, MC	31	34	0.81	0.069
		Kang et al., 2007	1	Passages	N	72	CR	MC	SA initial test	Read statements	Final MC	48		-0.30	0.023
			2	Passages	Y	72	CR	MC	SA initial test	Read statements	Final MC	48		0.47	0.024
		Avci, 2011	4	Passages	N	168	CR	MC	SA condition	Reread	Final MC	41		0.53	0.029
		Hinze & Wiley, 2011	3	Passages	N	48	CR	MC	Paragraph recall	Reread	Final test	25		0.44	0.048
			3	Passages	Y (R)	48	CR	MC	Paragraph recall	Reread	Final test	25		0.35	0.047
		McDaniel et al., 2012	1	Facts	Y (E)	504	CR	MC	Short answer	Read	Identical	16		1.05	0.12
			2	Facts	Y (E)	264	CR	MC	Short answer	Read	Identical	27		0.82	0.055
		Wooldridge et al., 2014	1	Passages	Y	48	CR	MC	Repeated fact	Highlight	Fact qs.	25	29	1.52	0.10
	Multiple-choice to cued recall														
		Nungester & Duchastel, 1982		Passages	N	336	MC	CR	Test	Review	Old items, SA	31	34	0.92	0.071
		Butler & Roediger, 2007		Video	Mixed**	672	MC	CR	Multiple choice	Study	Final test	27		0.01	0.040
		Kang et al., 2007	1	Passages	N	72	MC	CR	MC initial test	Read statements	Final SA	48		0.28	0.023
			2	Passages	Y	72	MC	CR	MC initial test	Read statements	Final SA	48		0.17	0.022
		Sensenig, 2010	1	Passages	N	0.083	MC	CR	Multiple choice	Re-studied	Final test	74		-0.44	0.015
			3	Passages	N	0.083	MC	CR	Multiple choice	Re-studied	Final test	54		0.18	0.020
		Avci, 2011	4	Passages	N	168	MC	CR	MC condition	Reread	Final SA	41		0.60	0.031
		Meyer & Logan, 2013		Facts	Y**	24	MC	CR	Tested‡	Restudied	Final test	180		1.03	0.0084
		Pan, Gopal, et al., 2015	3	Facts	Y	48	MC	CR	Tested	Restudied	Final test	52		0.52	0.023

(table continues)

Table 1 (*continued*)

				Study design											
Category	Sub-category	Reference	Exp.	Stimuli	Feedback	Delay (hr)	Train	Final	Initial test condition	Reexposure control	Transfer test	N_T	N_R	*d*	*sv*
Stimulus-response rearrangement															
	Paired associates														
		Carpenter et al., 2006	1	PAL	Y	33	CR	CR	Test/study	Study	? → B	53		0.86	0.026
			1	PAL	Y	33	CR	FR	Test/study	Study	Recall As	45		0.60	0.028
			2	PAL	Y	33	CR	CR	Test/study	Study	? → B	19		0.48	0.067
			2	PAL	Y	33	CR	FR	Test/study	Study	Recall As	18		1.01	0.10
		Cheng, 2014	1	PAL	N	0.25	MC	CR	Test	Restudy	Backward	50		0.72	0.027
		Rickard & Pan, 2018	1	PAL	Y	24	CR	CR	Tested	Restudied	Tested-rearranged	33		0.89	0.046
			1	PAL	Y	168	CR	CR	Tested	Restudied	Tested-rearranged	25		0.72	0.057
			App.	PAL	Y	24	CR	CR	Tested	Restudied	Tested-rearranged	33		0.58	0.038
	Triple associates														
		Pan, Wong, et al., 2016	1	Triplets	Y	168	CR	CR	Tested	Restudied	Tested-inverted	42		0.23	0.026
			2	Triplets	Y	168	CR	CR	Tested	Restudied	Tested-inverted	58		0.16	0.018
		Rickard & Pan, 2018	1	Triplets	Y	24	CR	CR	Tested	Restudied	Tested-rearranged	32		0.06	0.033
			1	Triplets	Y	168	CR	CR	Tested	Restudied	Tested-rearranged	29		0.08	0.037
			App.	Triplets	Y	24	CR	CR	Tested	Restudied	Tested-rearranged	35		0.08	0.030
	Multi-term facts														
		McDaniel et al., 2007		Chapters	Y(D)	504	CR	MC	Quizzed	Read only	Unit exam	34		0.41	0.038
		Hinze & Wiley, 2011	1	Passages	N	48	CR	CR	FITB testing	Restatement	Related	69		0.14	0.015
			2	Passages	N	168	CR	CR	FITB testing	Restatement	Related	26		-0.07	0.042
		Pan, Gopal, et al., 2015	1	Facts	Y	48	CR	CR	Tested	Restudied	Transfer	38		-0.01	0.028
			2	Facts	Y	48	CR	CR	Tested, 2x	Restudied, 2x	Transfer, 2x	58		-0.09	0.018
			3	Facts	Y	48	MC	CR	Tested	Restudied	Transfer	52		-0.30	0.021
			4	Facts	Y	24	CR	CR	Tested, 1x	Restudied, 1x	Transfer, 1x	45		-0.22	0.024
			4	Facts	Y	24	CR	CR	Tested, 2x	Restudied, 2x	Transfer, 2x	45		-0.01	0.023
		Pan, Hutter, et al., 2018	1	Facts	Y	48	CR	CR	Tested	Restudied	Transfer	51		0.03	0.020
			2	Facts	Y	48	CR	CR	Tested	Restudied	Transfer	57		-0.17	0.018
			3	Facts	Y	48	CR	CR	Tested	Restudied	Transfer	56		0.19	0.019
			4	Facts	Y	48	CR	CR	Tested	Restudied	Transfer	57		0.16	0.018
	Term-definition facts														
		McDaniel et al., 2015	1	Passage	N	120	MC	CR	TTT* (BE)	SSS*	Diff. stem, definition	26	25	0.87	0.090
			1	Passage	Y (R)	120	MC	CR	TST (BE)	SSS	Diff. stem, definition	24	25	0.39	0.087
			2	Passage	Y	120	MC	CR	TTT* (BE)	SSS*	Definition	36	35	0.10	0.058
			2	Passage	Y (R)	120	MC	CR	TST (BE)	SSS	Definition	36	35	0.59	0.061
		Pan & Rickard, 2017	2	Facts	Y	48	MC	MC	Tested	Restudied	Tested-different	31		0.21	0.035
			2	Facts	Y	48	MC	MC	Tested	Restudied	Tested-different	28		0.14	0.039
			3a	Facts	Y	48	MC	CR	Tested	Restudied	Tested-different	49		0.09	0.021
			3b	Facts	Y	48	MC	CR	Tested	Restudied	Tested-different	52		-0.02	0.020
			4	Facts	Y	48	MC	CR	Tested	Restudied	Tested-different	54		0.15	0.019
			4	Facts	Y	48	MC	CR	Tested	Restudied	Tested-different	59		0.09	0.018

(table continues)

Table 1 (*continued*)

				Study design										
Category	Sub-category	Reference	Exp.	Stimuli	Feedback	Delay (hr)	Train	Final	Initial test condition	Reexposure control	Transfer test	N_T	N_R	*d*	*sv*
Untested materials seen during initial study															
	None specified														
		La Porte & Voss, 1975		Passages	N	168	CR	CR	Question 50	Statement 50	New qs.	24	24	-0.18	0.088
				Passages	N	168	CR	CR	Question 100*	Statement 100*	New qs.	24	24	0.36	0.089
				Passages	Y	168	CR	CR	Question KR 50*	Statement 50*	New qs.	24	24	0.22	0.088
				Passages	Y	168	CR	CR	Question KR 100	Statement 100	New qs.	24	24	-0.09	0.087
		Nungester & Duchastel, 1982		Passage	N	336	MC/CR	MC/CR	Test	Review	Subset B	31	34	-0.25	0.064
		Balch, 1998	1	Course	Y (R)	168	MC	MC	Practice-exam	Review	Final test	66	66	0.37	0.031
		Chan et al., 2006	1	Passage	N	24	CR	CR	Testing	Extra study	Not presented, Day 1	28	28	0.70	0.079
			3	Passages	N	24	CR	CR	Narrow retrieval*	Extra study*	Related	18	18	0.46	0.14
			3	Passages	N	24	CR	CR	Broad retrieval (BE)	Extra study	Related	18	18	1.01	0.13
		Pilotti et al., 2009		Course	Y	0	MC	MC	RP	Reviewing	Final test	127	129	-0.13	0.016
		Butler, 2010	1a/1b/2	Passages	Y	168	CR	CR	Same/variable test‡	Re-study passages	Control	72		-0.44	0.016
		Agarwal, 2011	1	Passages	Y	48	MC	MC	Higher order quiz	Study twice	Concept qs.	48		-0.03	0.022
		Little, 2011	5	Passages	Y (R)	0.083	MC	CR	Standard	Extended-study	Related	28		-0.13	0.039
			5	Passages	Y (R)	0.083	MC	CR	Discrimination (BE)	Extended-study	Related	28		0.41	0.042
		McDaniel et al., 2012	1	Chapters	Y (E)	504	MC	MC	Multiple-choice*	Read*	Related	16		0.48	0.082
			1	Chapters	Y (E)	504	CR	MC	Short answer	Read	Related	16		0.49	0.083
		Hinze et al., 2013	2	Passages	N	192	CR	MC	Paragraph recall	Reread	Detail test	25	26	0.82	0.089
			2	Passages	N	192	CR	MC	Expect detail*	Reread*	Detail test	23	26	0.06	0.086
			2	Passages	N	192	CR	MC	Expect inference*	Reread*	Detail test	23	26	0.58	0.090
			3	Passages	N	192	FR	MC	Free recall*	Reread*	Detail test	31	31	0.16	0.067
			3	Passages	N	192	FR	MC	Explain (BE)	Reread	Detail test	31	31	0.65	0.071
		Hinze & Rapp, 2014	1	Passages	N	168	CR	MC	Low-stakes quiz	Rereading	Detail	21	21	-0.22	0.10
			1	Passages	N	168	CR	NC	High-stakes quiz*	Rereading*	Detail	19	21	-0.84	0.12
		Wooldridge et al., 2014	1	Chapter	Y	48	CR	MC	Related fact quiz	Highlight	Fact qs.	20	29	0.31	0.088
			2	Chapter	Y (R)	48	CR	MC	Quiz-restudy	Highlight	Fact qs.	24	24	0.14	0.087
Application and inference															
	Application questions only														
		Hinze & Rapp, 2014	1	Passages	N	168	CR	CR	Low-stakes quiz	Rereading	Application	21	21	0.65	0.11
			1	Passages	N	168	CR	CR	High-stakes quiz*	Rereading*	Application	19	21	-0.08	0.11
		Wooldridge et al., 2014	1	Chapter	Y	48	CR	MC	Related application	Highlight	Application	30	24	-0.33	0.079
			2	Chapter	Y (R)	48	CR	MC	Quiz-restudy	Highlight	Application	24	24	0.12	0.087
		McDaniel et al., 2015	1	Passage	N	120	MC	CR	TTT*(BE)	SSS*	Diff. stem, application	26	25	0.48	0.084
			1	Passage	Y (R)	120	MC	CR	TST (BE)	SSS	Diff. stem, application	24	25	0.59	0.089
			2	Passage	Y	120	MC	CR	TTT*(BE)	SSS*	Application	36	35	0.29	0.059
			2	Passage	Y (R)	120	MC	CR	TST (BE)	SSS	Application	36	35	0.68	0.062
		Nguyen & McDaniel, 2016	1	Passages	Y (R)	0.37	FR	CR	Standard 3R	Meta-notetaking	Problem-solving	32	32	0.14	0.065

(table continues)

Table 1 (*continued*)

				Study design											
Category	Sub-category	Reference	Exp.	Stimuli	Feedback	Delay (hr)	Train	Final	Initial test condition	Reexposure control	Transfer test	N_T	N_R	*d*	*sv*
			1	Passages	Y (R)	0.37	FR	CR	Meta-3R*	Meta-notetaking*	Prob.-solving	32	32	0.14	0.065
		Nguyen et al., 2016	2	Passages	Y	48	MC	MC	Novel quiz	Novel presentation	Transfer	22	23	-0.02	0.093
			2	Passages	Y	48	MC	MC	Text-verbatim quiz*	Novel presentation*	Final test	22	23	-0.61	0.098
	Inference questions only														
		Hinze & Rapp, 2014	1	Passages	N	168	CR	MC	Low-stakes quiz	Rereading	Inference	21	21	0.53	0.10
			1	Passages	N	168	CR	MC	High-stakes quiz*	Rereading*	Inference	19	21	-0.30	0.11
		Tran et al., 2014	1	Premises	Y	0	CR	MC	RP	Reread	Final test	68		0.16	0.015
			2	Premises	Y	48	CR	MC	RP	Reread	Final test	40		-0.06	0.026
			3	Premises	Y	0	CR	MC	RP	Reread	Final test	84		-0.21	0.012
			3	Premises	Y	48	CR	MC	RP	Reread	Final test	84		-0.05	0.012
			4	Premises	Y	48	CR	MC	RP	Reread	Inference	84		-0.19	0.012
		Lechuga et al., 2015		Passage	Y (R)	168	FR	CR	Repeated retrieval	Rpt. study	Inference	20	20	0.15	0.11
		Eglington & Kang, 2016	App.	Premises	Y	49.5	CR	MC	RP	Restudy	Final test	45		0.07	0.023
			1	Premises	Y(D)	49.5	CR	MC	RP	Restudy	Final test	56		0.38	0.020
			2	Premises	Y(D)	49.5	CR	MC	RP, simultaneous	Restudy	Final test	64		0.35	0.017
			2	Premises	Y	49.5	CR	MC	RP, single	Restudy	Final test	64		-0.09	0.016
			3	Premises	Y(D)	49.5	FR	MC	RP	Restudy	Final test	45		0.38	0.025
	Application and inference questions														
		Johnson & Mayer, 2009		Anim.	N	0.083	CR	CR	Practice-retention*	Restudy*	New qs.	53	53	-0.46	0.040
				Anim.	N	168	CR	CR	Practice-retention*	Restudy*	New qs.	59	59	-0.23	0.035
				Anim.	N	0.083	CR	CR	Practice-transfer	Restudy	New qs.	27	53	0.19	0.058
				Anim.	N	168	CR	CR	Practice-transfer	Restudy	New qs.	31	59	0.68	0.078
		McDaniel et al., 2009	1	Passages	Y (R)	0	FR	CR	3R	Rereading	Inference	24	24	-0.07	0.087
			1	Passages	Y (R)	168	FR	CR	3R	Rereading	Inference	24	24	0.26	0.088
			2	Passages	Y (R)	0	FR	CR	3R	Rereading	Prob.-solving	24	24	0.65	0.092
			2	Passages	Y (R)	168	FR	CR	3R	Rereading	Prob.-solving	24	24	0.65	0.092
		Butler, 2010	1b	Passages	Y	168	CR	CR	Same test	Re-study passages	Factual inf.	24		1.17	0.081
			1b	Passages	Y	168	CR	CR	Variable test*	Re-study passages*	Factual inf,	24		0.89	0.066
			1b	Passages	Y (E)	168	CR	CR	Same test*	Re-study passages*	Conceptual inf.	24		0.91	0.067
			1b	Passages	Y (E)	168	CR	CR	Variable test	Re-study passages	Conceptual inf.	24		1.04	0.074
			2	Passages	Y	168	CR	CR	Same test*	Re-study sentences*	Factual inf.	24		0.66	0.057
			2	Passages	Y (E)	168	CR	CR	Variable test	Re-study sentences	Conceptual inf.	24		0.53	0.053
			3	Passages	Y (E)	168	CR	CR	Same test	Re-study passages	Conceptual inf.	20		1.17	0.10
		Agarwal, 2011	1	Passages	Y	48	MC	MC	Concept quiz	Study twice	Higher order	48		-0.11	0.022
		Karpicke & Blunt,2011	1	Passage	Y (R)	168	FR	CR	RP	Rpt. study	Inference qs.	20	20	0.50	0.11
		Hinze et al., 2013	2	Passages	N	168	CR	MC	Paragraph recall	Reread	Inference	25	26	0.53	0.085
			2	Passages	N	168	CR	MC	Expect detail*	Reread*	Inference	23	26	0.10	0.086
			2	Passages	N	168	CR	MC	Expect inference*	Reread*	Inference	23	26	0.68	0.091
			3	Passages	N	168	FR	MC	Free recall*	Reread*	Inference	31	31	0.19	0.067

(table continues)

Table 1 (*continued*)

				Study design											
Category	Sub-category	Reference	Exp.	Stimuli	Feedback	Delay (hr)	Train	Final	Initial test condition	Reexposure control	Transfer test	N_T	N_R	*d*	*sv*
			3	Passages	N	168	CR	MC	Explain (BE)	Reread	Inference	31	31	0.49	0.069
		Zhou et al., 2013	1	Passage	Y (R)	168	FR	CR	RP	Restudy	Inference	40	40	0.22	0.052
			2	Passage	Y (R)	168	FR	CR	RP	Concept mapping	Inference	43		0.32	0.026
		Blunt & Karpicke, 2014	2	Passage	Y (R)	168	FR	CR	RP, paragraph	Rpt. Study, paragraph	Inference	20	20	0.58	0.11
			2	Passage	Y (R)	168	FR	CR	RP, concept map	Rpt. Study, concept map	Inference	20	20	0.35	0.11
		Nguyen & McDaniel, 2016	1	Passages	Y (R)	0.37	FR	MC	Standard 3R	Meta-notetaking	Inference	32	32	-0.27	0.065
			1	Passages	Y (R)	0.37	FR	MC	Meta-3R*	Meta-notetaking*	Inference	32	32	0.46	0.066
		van Eersel et al., 2016	1	Passages	Y (E)	0.083	CR	CR	Testing	Rereading	Final test	24		1.23	0.085
			1	Passages	Y (E)	168	CR	CR	Testing	Rereading	Final test	24		1.08	0.076
			2	Passages	Y (E)	168	CR	CR	Testing	Reread-plus-feedback	Final test	54		0.42	0.021
Problem-solving skills															
	Medical diagnosis and treatment														
		Kromann et al., 2009		Course	Y**	336	CST	CST	Intervention	Control	Final test	41	40	0.93	0.056
		Kromann et al., 2010		Course	Y**	4320	CST	CST	Intervention	Control	Final test	48	41	0.40	0.047
		Kromann et al., 2011		Course	Y**	336	CST	CST	Intervention	Control	Final test	66	72	0.55	0.031
		Larsen et al., 2013a		Course	Y (D)	4320	SA	CR	Test, no self-expl.	Study, no self-expl.	Final test	49		0.70	0.027
				Course	Y (D)	4320	SA	CR	Test, self-expl. (BE)	Study, self-expl.	Final test	49		0.48	0.024
		Larsen et al., 2013b		Proced.	Y (D)	4320	SPT	SPT	Standardized patient test	Review	SPT	41		0.80	0.034
				Proced.	Y (D)	4320	SA	SPT	Written test*	Review*	SPT	41		0.33	0.027
	Worked examples														
		Darabi et al., 2007		Probs.	N	342	PST	PST	Problem group	Product group	Transfer test	22	22	0.96	0.11
		van Gog et al., 2011		Probs.	N	0	PST	PST	Example-problem	Example-example	Final test	22	26	-0.02	0.088
		van Gog & Kester, 2012		Probs.	N	0.083	PST	PST	STST-T-T	SSSS-T-T	Final test	20	20	-0.25	0.11
		Leahy et al., 2015	1	Probs.	Y	0	PST	PST	Worked examples+prob.	Worked examples only	Final test	17	16	-0.73	0.14
			2	Probs.	Y	0	PST	PST	Worked examples+prob.	Worked examples only	Final test	24	25	-0.68	0.091
			3	Probs.	Y	168	PST	PST	Worked examples+prob.	Worked examples only	Final test	14	13	0.05	0.16
		van Gog et al., 2015	1	Probs.	N	0.083	PST	PST	EPEP, isomorphic	EEEE, isomorphic	Final test	20	20	0.24	0.11
			1	Probs.	N	168	PST	PST	EPEP, isomorphic*	EEEE, isomorphic*	Final test	20	20	0.04	0.11
			1	Probs.	N	0.083	FR	PST	ERER, isomorphic*	EEEE, isomorphic*	Final test	20	20	-0.10	0.11
			1	Probs.	N	168	FR	PST	ERER, isomorphic	EEEE, isomorphic	Final test	20	20	0.22	0.11
			2	Probs.	N	0.083	PST	PST	EPEP	EEEE	Isomorphic	15	15	0.21	0.14
			2	Probs.	N	168	PST	PST	EPEP*	EEEE*	Isomorphic	15	15	-0.08	0.14
			3	Probs.	N	168	PST	PST	EEEE-PP-isomorphic	EEEE-EE-isomorphic	Final test	29	34	0.07	0.066
			3	Probs.	N	168	PST	PST	EEEE-PP-identical	EEEE-EE-identical	Final test	33	33	0.04	0.063
			4	Probs.	N	168	PST	PST	Testing	Restudy	Final test	30	27	0.34	0.074

(table continues)

Table 1 (*continued*)

				Study design											
Category	Sub-category	Reference	Exp.	Stimuli	Feedback	Delay (hr)	Train	Final	Initial test condition	Reexposure control	Transfer test	N_T	N_R	*d*	*sv*
Mediator and related word cues															
	Mediator cues														
		Carpenter, 2011	2	PAL	N	0.5	CR	CR	Test	Study	Semantic mediator	10	10	1.84	0.34
		Rawson et al., 2015	1	PAL	N	48	CR	CR	Test	Restudy	Mediator, short lag	33	31	1.19	0.077
			1	PAL	N	48	CR	CR	Test	Restudy	Mediator, long lag	31	30	1.25	0.082
			2	PAL	N	48	CR	CR	STSTST	Restudy	Mediator, short lag	50	40	0.67	0.049
			2	PAL	N	48	CR	CR	STSTST*	Restudy*	Mediator, long lag	53	45	1.06	0.048
			2	PAL	N	48	CR	CR	SSSTTT*	Restudy*	Mediator, short lag	45	40	0.38	0.049
			2	PAL	N	48	CR	CR	SSSTTT	Restudy	Mediator, long lag	51	45	0.95	0.048
			App.	PAL	N	0.33	CR	CR	STT	Restudy	Mediator, short lag	32	32	0.71	0.069
			App.	PAL	N	0.33	CR	CR	STT	Restudy	Mediator, long lag	32	32	0.57	0.067
		Coppens et al., 2016	1	PAL	N	0.5	CR	CR	Testing, no-mc	Restudy	Mediator cues	17	28	0.52	0.10
			1	PAL	N	0.5	CR	CR	Testing, strong-mc	Restudy	Mediator cues	23	26	0.24	0.086
			2	PAL	N	0.5	CR	CR	Testing	Restudy	Mediator cues	47	44	0.42	0.046
			3	PAL	N	0.5	CR	CR	Testing	Restudy	Mediator cues	26	30	0.94	0.083
		Cho et al., 2017	1	PAL	N	0.5	CR	CR	Test	Restudy	.20 M-C Pairs, mediator	38	38	0.50	0.056
			1	PAL	N	0.5	CR	CR	Test	Restudy	.00 M-C Pairs, mediator	38	38	0.89	0.060
	Related cues														
		Carpenter, 2011	2	PAL	N	0.5	CR	CR	Test	Study	New related	10	10	0.65	0.24
		Veltre et al., 2013		Words	Y	48	CR	CR	Semantic/different-type	Restudy	Final test	48	48	0.54	0.044
		Rawson et al., 2015	1	PAL	N	48	CR	CR	Test	Restudy	Related, short lag	33	31	0.42	0.066
			1	PAL	N	48	CR	CR	Test	Restudy	Related, long lag	31	30	1.18	0.08
			2	PAL	N	48	CR	CR	STSTST	Restudy	Related, short lag	50	40	0.48	0.047
			2	PAL	N	48	CR	CR	STSTST*	Restudy*	Related, long lag	53	45	0.32	0.043
			2	PAL	N	48	CR	CR	SSSTTT*	Restudy*	Related, short lag	45	40	0.26	0.049
			2	PAL	N	48	CR	CR	SSSTTT	Restudy	Related, long lag	51	45	0.57	0.045
			App.	PAL	N	0.33	CR	CR	STT	Restudy	Related, short lag	32	32	0.37	0.066
			App.	PAL	N	0.33	CR	CR	STT	Restudy	Related, long lag	32	32	0.43	0.066
		Coppens et al., 2016	1	PAL	N	0.5	CR	CR	Testing, no-mc	Restudy	Related cues	21	22	0.26	0.099
			1	PAL	N	0.5	CR	CR	Testing, strong-mc	Restudy	Related cues	30	21	1.23	0.10
			2	PAL	N	0.5	CR	CR	Testing	Restudy	Related cues	36	38	0.49	0.057
			3	PAL	N	0.5	CR	CR	Testing	Restudy	Related cues	31	29	-0.04	0.069
		Cho et al., 2017	1	PAL	N	0.5	CR	CR	Test	Restudy	.20 M-C Pairs, related	41	41	0.32	0.051
			1	PAL	N	0.5	CR	CR	Test	Restudy	.00 M-C Pairs, related	41	41	0.08	0.050

Note. Under Stimuli, PAL = Paired associates; CAT = Categories. Under Feedback (Correct Answer Feedback), Y = Yes and N = No; (R) = Restudy; (D) = Extended and/or detailed feedback; (E) = Explanatory feedback. Under Train and Final (Test format), FR = Free recall, CR = Cued recall, MC = Multiple-choice, Recog. = Recognition, CST = Clinical scenario test, SPT = Simulated patient test, PST = Problem-solving test. RP = Retrieval practice, Proced. = Procedures, Probs. = Problems, qs. = Questions, Self-expl, = Self-explanations. Under Initial test condition, (BE) = Broad encoding methods. Other abbreviations are drawn verbatim from, and defined in, the original articles. (*) indicates data that was randomly excluded from the meta-analyses due to non-independence with other included conditions; (**) indicates feedback was provided but did not qualify as correct answer feedback; (‡) indicates data reported collapsed across conditions. Experiment numbers of unpublished studies may differ from that in subsequently published articles.

Table 2.

Overall and Category-Level Weighted Mean Effect Sizes

Category and dataset (number of effect sizes)	*β*	*SE*	*df*	*p*	95% C.I.
Overall across all categories (*k* = 192)	0.40	0.046	43.016	<.00001	[0.31,0.50]
Initial test performance data available (*k* = 135)	0.41	0.054	29.01	<.00001	[0.30,0.52]
Test format (*k* = 56)	0.58	0.071	20.86	<.00001	[0.43,0.73]
Stimulus-response rearrangement (*k* = 33)	0.22	0.098	6.46	.066	[-0.019,0.45]
Untested materials seen during initial study (*k* = 17)	0.16	0.12	10.086	.20	[-0.10,0.43]
Application and inference questions (*k* = 41)	0.33	0.11	10.39	.013	[0.085,0.56]
Problem-solving skills (*k* = 18)	0.29	0.15	5.46	.10	[-0.078,0.65]
Mediator and related word cues (*k* = 27)	0.61	0.089	2.12	.018	[0.25,0.97]

Note. *β* = regression coefficient in terms of Cohen's *d*; *SE* = standard error; *df* = adjusted degrees of freedom, C.I. = confidence interval.

Table 3.

Overall Random-Effects Meta-Analyses Results

Analysis type	Dataset	Model type	Moderator variable/intercept	*β*	*SE*	*df*	*p*	95% C.I.
Single moderator fits								
	All effect sizes (k = 192)							
		–	Response congruency	0.30	0.081	35.55	.0006	[0.14,0.47]
	Initial test performance data available (k = 135)							
		–	Initial test performance	0.82	0.16	15.30	.0001	[0.47,1.17]
			No. training repetitions	0.13	0.032	13.75	.0011	[0.06,0.20]
			Response congruency	0.29	0.094	26.68	.0048	[0.096,0.48]
Simultaneous moderator fits								
	All effect sizes (k = 192)							
		Main effect of all moderators, except initial test performance						
			Elaborated retrieval practice	0.22	0.077	23.56	.0094	[0.059, 0.38]
			Response congruency	0.35	0.082	35.37	.0002	[0.18, 0.51]
			Intercept	0.21	0.064	21.11	.0031	[0.081, 0.35]
	Initial test performance data available (k = 135)							
		Main effect of all moderators, including initial test performance						
			Initial test performance	0.58	0.17	15.96	.0029	[0.23, 0.93]
			Elaborated retrieval practice	0.23	0.085	17.91	.015	[0.050, 0.41]
			Response congruency	0.26	0.086	23.08	.0058	[0.084, 0.44]
			Intercept	-0.16	0.10	12.65	.13	[-0.38, 0.58]

Note. *β* = regression coefficient in terms of Cohen's *d*; *SE* = standard error; *df* = adjusted degrees of freedom, C.I. = confidence interval, No. = number. The intercept is reported for all simultaneous moderator fits.

Table 4.

Overall PEESE Analyses Results

Analysis type	Dataset	Moderator variable/intercept	β	t	p
No moderators fitted					
	All effect sizes (k = 192)				
		Sampling variability	4.41	4.51	<.0001
		Intercept	0.17	3.69	.0003
	Initial test performance data available (k = 135)				
		Sampling variability	5.61	4.48	<.0001
		Intercept	0.013	2.41	.017
With moderators fitted					
	All effect sizes (k = 192)				
		Sampling variability	3.86	4.32	<.0001
		Elaborated retrieval practice	0.18	2.65	.0088
		Response congruency	0.36	6.89	<.0001
		Intercept	0.015	0.33	.74
	Initial test performance data available (k = 135)				
		Sampling variability	4.53	3.89	.0002
		Initial test performance	0.50	3.39	.0009
		Elaborated retrieval practice	0.14	1.64	.10
		Response congruency	0.25	4.16	<.0001
		Intercept	-0.30	-2.89	.0045

Note. β = regression coefficient in terms of Cohen's d.

Table 5.

Overall Effect Size Estimates for Various Publication Bias Scenarios

			Publication Bias Scenario			
Dataset	Moderator variable/intercept	Unadjusted estimate	Moderate one-tailed	Severe one-tailed	Moderate two-tailed	Severe two-tailed
All effect sizes (k = 192)						
	Elaborated retrieval practice	0.22	0.22	0.27	0.20	0.17
	Response congruency	0.35	0.37	0.48	0.33	0.31
	Intercept	0.21	0.12	-0.12	0.17	0.12
Initial test performance data available (k = 135)						
	Initial test performance	0.58	0.60	0.70	0.53	0.46
	Elaborated retrieval practice	0.23	0.24	0.27	0.21	0.19
	Response congruency	0.26	0.28	0.35	0.25	0.24
	Intercept	-0.16	-0.26	-0.53	-0.17	-0.17

Note. Effect size estimates are in terms of Cohen's *d*. All effect size estimates were derived using the selection methods detailed in Vevea and Woods (2005) and with the suggested *p*-value cutoffs of .001, .01, .05, and .50 (and for two-tailed selection methods, also .95, .99, and .999). Investigated moderators followed that of the random-effects meta-analyses.

Table 6.

Category-Level Random-Effects Meta-Analyses Results

Category	Analysis type	Dataset	Moderator variable/intercept	*β*	*SE*	*df*	*p*	95% C.I.
Test format								
	Single moderator fits							
		All effect sizes (k = 56)	Between- vs. within-subjects design	0.35	0.11	12.63	.0062	[0.12, 0.58]
	Simultaneous moderator fits							
		All effect sizes (k = 56)	Between- vs. within-subjects design	0.35	0.11	12.63	.0062	[0.12, 0.58]
			Intercept	0.49	0.077	15.20	<.0001	[0.33, 0.66]
		Initial test performance data available (k = 46)	Between- vs. within-subjects design	0.40	0.18	9.23	.048	[0.42, 0.80]
			Initial test performance	0.86	0.29	9.25	.016	[0.20, 1.51]
			Multiple-choice vs. not on the initial test	-0.48	0.16	5.73	.024	[-0.88, -0.093]
			Response congruency	0.29	0.11	6.53	.039	[0.019, 0.56]
			Intercept	-0.27	0.25	7.10	.33	[-0.87, 0.33]
Stimulus-response rearrangement								
	Single moderator fits							
		All effect sizes (k = 33)	Paired associates vs. non-paired associates	0.66	0.053	2.41	.0031*	[0.46, 0.86]
	Simultaneous moderator fits							
		All effect sizes (k = 33)	Paired associates vs. non-paired associates	0.66	0.053	2.41	.0031*	[0.46, 0.86]
			Intercept	0.063	0.054	4.62	.29	[-0.078, 0.20]
Untested materials seen during initial study								
	Single moderator fits							
		All effect sizes (k = 17)	Elaborated retrieval practice	0.37	0.14	7.87	.032	[0.041, 0.70]
	Simultaneous moderator fits							
		All effect sizes (k = 17)	Elaborated retrieval practice	0.37	0.14	7.87	.032	[0.041, 0.70]
			Intercept	0.0028	0.13	7.31	.98	[-0.29, 0.29]

(table continues)

Table 6. (*continued*)

Category	Analysis type	Dataset	Moderator variable/intercept	*β*	*SE*	*df*	*p*	95% C.I.
Application and inference questions								
	Single moderator fits							
		All effect sizes ($k = 41$)	No. training repetitions	0.33	0.072	3.31	.016*	[0.11,0.55]
			Retention interval	0.0033	0.0009	9.64	.0060	[0.0012,0.0054]
			Elaborated retrieval practice	0.35	0.13	7.97	.029	[0.046,0.66]
	Simultaneous moderator fits							
		All effect sizes ($k = 41$)	Correct answer feedback	-0.49	0.074	2.62	.011*	[-0.75, -0.24]
			No. training repetitions	0.29	0.098	3.27	.054*	[-0.095, 0.58]
			Elaborated retrieval practice	0.26	0.11	5.54	.063	[-0.019, 0.54]
			Intercept	0.44	0.062	2.02	.019	[0.18, 0.71]
Problem-solving skills								
	Single moderator fits							
		All effect sizes ($k = 18$)	Worked examples vs. medical diagnosis and treatment	0.59	0.20	5.34	.028	[0.093, 1.09]
	Simultaneous moderator fits							
		All effect sizes ($k = 18$)	Worked examples vs. medical diagnosis and treatment	0.59	0.20	5.34	.028	[0.093, 1.09]
			Intercept	0.45	0.18	2.17	.83	[-0.69, 0.78]

Note. β = regression coefficient in terms of Cohen's *d*; *SE* = standard error; *df* = adjusted degrees of freedom, C.I. = confidence interval, No. = number. An asterisk indicates that the *p*-value may be untrustworthy due to insufficient degrees of freedom (< 4). The intercept is reported for all simultaneous moderator fits. No single or simultaneous moderator fits were performed for the category of transfer to mediator and related word cues, owing to too few studies currently available in that category.

Table 7.

Category-Level PEESE Analyses Results

Analysis type	Category	Dataset	Moderator variable/intercept	β	t	p
No moderators fitted						
	Test format					
		All effect sizes (k = 56)	Sampling variability	5.06	2.35	.023
			Intercept	0.36	4.14	<.0001
		Initial test performance data available (k = 46)	Sampling variability	6.74	2.71	.0096
			Intercept	0.25	2.71	.0096
	Stimulus-response rearrangement					
		All effect sizes (k = 33)	Sampling variability	10.93	3.30	.0025
			Intercept	-0.12	-1.23	.2265
	Untested materials seen during initial study					
		All effect sizes (k = 17)	Sampling variability	6.91	2.75	.0148
			Intercept	-0.26	-1.91	.076
	Application and inference questions					
		All effect sizes (k = 41)	Sampling variability	6.43	4.27	<.0001
			Intercept	-0.045	-0.61	.55
	Problem-solving skills					
		All effect sizes (k = 18)	Sampling variability	-6.56	-2.92	.010
			Intercept	0.78	4.97	<.0001
	Mediator and related word cues					
		All effect sizes (k = 27)	Sampling variability	3.55	1.90	.070
			Intercept	0.36	2.58	.016
With moderators fitted						
	Test format					
		All effect sizes (k = 56)	Sampling variability	3.19	1.07	.29
			Between- vs. within-subjects design	0.20	0.90	.37
			Intercept	0.39	4.16	<.0001
		Initial test performance data available (k = 46)	Sampling variability	6.23	2.33	.025
			Between- vs. within-subjects design	0.14	0.63	.53
			Initial test performance	0.97	3.45	.0013
			Multiple-choice vs. not on the initial test	-0.44	-3.46	.0013
			Response congruency	0.26	2.06	.046
			Intercept	-0.55	-2.04	.048
	Stimulus-response rearrangement					
		All effect sizes (k = 33)	Sampling variability	4.35	1.93	.064
			Paired associates vs. non-paire associates	0.59	7.1	<.0001
			Intercept	-0.044	-0.70	.49

(table continues)

Table 7. (*continued*)

Analysis type	Category	Dataset	Moderator variable/intercept	*β*	*t*	*p*
	Untested materials seen during initial study					
		All effect sizes (*k* = 17)	Sampling variability	5.68	2.46	.028
			Elaborated retrieval practice	0.34	2.18	.047
			Intercept	-0.32	-2.54	.023
	Application and inference questions					
		All effect sizes (*k* = 41)	Sampling variability	1.94	1.34	.11
			Correct answer feedback	-0.38	-2.37	.024
			No. training repetitions	0.23	3.66	.0008
			Elaborated retrieval practice	0.26	2.98	.0051
			Intercept	0.28	1.63	.11
	Problem-solving skills					
		All effect sizes (*k* = 18)	Sampling variability	-0.039	-0.01	.99
			Worked examples vs. medical diagnosis and treatment	0.57	2.02	.061
			Intercept	0.051	0.13	.90

Note. *β* = regression coefficient in terms of Cohen's *d*. The order of the category-level PEESE analyses, number of analyses, and the moderators investigated in these analyses, followed those used in the random-effects meta-analyses.

Table 8.

Category-Level Effect Size Estimates for Various Publication Bias Scenarios

				Publication Bias Scenario			
Category	Dataset	Moderator variable/intercept	Unadjusted estimate	Moderate one-tailed	Severe one-tailed	Moderate two-tailed	Severe two-tailed
Test format							
	All effect sizes (k = 56)	Between- vs. within-subjects design	0.35	0.36	0.43	0.34	0.31
		Intercept	0.49	0.41	0.20	0.44	0.37
	Initial test performance data available (k = 46)	Between- vs. within-subjects design	0.40	0.42	0.46	0.40	0.38
		Initial test performance	0.86	0.98	1.20	0.89	0.91
		Multiple-choice vs. not on the initial test	-0.48	-0.54	-0.66	-0.48	-0.46
		Response congruency	0.29	0.33	0.39	0.30	0.30
		Intercept	-0.28	-0.45	-0.75	-0.34	-0.41
Stimulus-response rearrangement							
	All effect sizes (k = 33)	Paired associates vs. non-paired associates	0.66	0.68	0.74	0.66	0.67
		Intercept	0.062	0.036	-0.036	0.052	0.036
Untested materials seen during initial study							
	All effect sizes (k = 17)	Elaborated retrieval practice	0.38	0.37	0.39	0.32	0.25
		Intercept	-0.0041	-0.071	-0.21	-0.014	-0.025
Application and inference questions							
	All effect sizes (k = 41)	Correct answer feedback	-0.49	-0.46	-0.45	-0.43	-0.35
		No. training repetitions	0.27	0.27	0.29	0.26	0.24
		Elaborated retrieval practice	0.28	0.26	0.25	0.25	0.20
		Intercept	0.43	0.38	0.30	0.39	0.31
Problem-solving skills							
	All effect sizes (k = 18)	Worked examples vs. medical diagnosis and treatment	0.58	0.62	0.77	0.58	0.58
		Intercept	0.047	-0.0049	-0.17	0.039	0.027
Mediator and related word cues	All effect sizes (k = 27)	Intercept	0.61	0.55	0.43	0.55	0.46

Note. Effect size estimates are in terms of Cohen's *d*. All effect size estimates were derived using the selection methods detailed in Vevea and Woods (2005) and with the suggested *p*-value cutoffs of .001, .01, .05, and .50 (and for two-tailed selection methods, also .95, .99, and .999). Investigated moderators followed that of the random-effects meta-analyses.