

**Iowa State University**

---

**From the Selected Works of Jason C.K. Chan**

---

2018

# Retrieval potentiates new learning: A theoretical and meta-analytic review

Jason C.K. Chan

Christian A Meissner

Sara D. Davis, *Iowa State University*



Available at: [https://works.bepress.com/jason\\_chan/30/](https://works.bepress.com/jason_chan/30/)

Retrieval Potentiates New Learning: A Theoretical and Meta-Analytic Review

Jason C.K. Chan

Christian A. Meissner

Sara D. Davis

*Iowa State University*

**Article published in *Psychological Bulletin***

This Copy is Before Copy Editing, Please Do Not Quote.

Address correspondence to:

Jason C.K. Chan

[ckchan@iastate.edu](mailto:ckchan@iastate.edu)

W112 Lagomarcino Hall  
Iowa State University  
Ames, IA 50011  
USA

### **Abstract**

A growing body of research has shown that retrieval can enhance future learning of new materials. In the present report, we provide a comprehensive review of the literature on this finding, which we term test-potentiated new learning. Our primary objectives were to: 1) produce an integrative review of the existing theoretical explanations, 2) summarize the extant empirical data with a meta-analysis, 3) evaluate the existing accounts with the meta-analytic results, and 4) highlight areas that deserve further investigations. Here, we identified four non-exclusive classes of theoretical accounts, including resource accounts, metacognitive accounts, context accounts, and integration accounts. Our quantitative review of the literature showed that testing reliably potentiates the future learning of new materials by increasing correct recall or by reducing erroneous intrusions, and several factors have a powerful impact on whether testing potentiates or impairs new learning. Results of a meta-regression analysis provide considerable support for the integration account. Lastly, we discuss areas of under-investigation and possible directions for future research.

### **Public Significance Statement**

Taking a test can enhance students' ability to learn new information later. Here, we provide a theoretical and quantitative synthesis of this literature. Our results show that testing enhances correct recall associated with new learning and reduces incorrect intrusions, but they also show that interrupting learning with tests too frequently can impair new learning.

**Keywords:** testing effect; learning and memory; test-potentiated learning; retrieval practice, forward effect of testing

### **Test-Potentiated New Learning: A Meta-Analytic Review**

The importance of retrieval processes was largely ignored in behavioral research of human memory for the better part of its first century, which is perhaps most evident in the relatively unknown status of Semon's pioneering work on retrieval until the late 1970s (Schacter, 2001; Schacter, Eich, & Tulving, 1978). However, following the seminal work by Tulving and his colleagues (Tulving & Pearlstone, 1966; Tulving & Thompson, 1973), students of memory have paid increasing attention to the importance of retrieval. Whereas early research primarily focused on uncovering the effects of different *retrieval conditions* on memory (e.g., the match between encoding and retrieval condition or the effects of retrieval intention on explicit and implicit memory), the last decade has seen a surge of interest in the effects of engaging in retrieval on later memory. In the present paper, we focus on the phenomenon of *test-potentiated new learning*, whereby attempting retrieval can enhance subsequent learning of new material. In the following, we present an integrative review of this literature. First, we provide a brief overview of the phenomenon and its historical origins. Second, we review the various explanations that have been proposed to account for the beneficial effects of retrieval on new learning, and we organize these explanations into four classes of theories. Third, we provide a quantitative summary of this literature in a meta-analysis. Fourth, we use a qualitative assessment and a meta-regression to evaluate the four theories. Lastly, we consider areas of research that require further investigation.

### **Retrieval Enhances Subsequent Learning of New Information**

The finding that retrieval potentiates subsequent learning is well established. Verbal learning studies have long reported that testing can strengthen subsequent relearning of the studied materials (Donaldson, 1971; Izawa, 1967; 1971; Young, 1971), a finding that has

generally been termed test-potentiated learning. More recent work has revealed that retrieval can also facilitate later learning of *new* information (Cho, Neely, Crocco, & Vitrano, 2017; Pierce & Hawthorne, 2016; Szpunar, McDermott, & Roediger, 2008). Researchers have studied the influence of retrieval on new learning extensively across a variety of study materials (e.g., word lists, paired associates, prose passages, trivia facts, slides, videos), test formats (e.g., free recall, cued recall, recognition), and subject populations (brain damaged patients, Pastötter, Weber, & Bauml, 2013; college students, Szpunar et al., 2008; e.g., children, Aslan & Bauml, 2015; online samples, Weinstein, Gilmore, Szpunar, & McDermott, 2014).

The future-oriented benefit of retrieval has profound implications for education, particularly for situations in which learners must sustain attention for an extended period of time, such as a lecture, a workshop, a webinar, etc. It is well known that students have difficulty sustaining attention throughout a lecture, and these lapses of attention often manifest as mind wandering, which are detrimental to learning (Smallwood & Schooler, 2015; Szpunar, 2017). Indeed, the frequency of mind wandering tends to rise throughout the duration of a lecture (Johnstone & Percival, 1976; Stuart & Rutherford, 1978), which suggests that students might have particular difficulty learning the materials presented in the later parts of a lecture. Given the continued importance of lecture-based instructions in global education, this time-related decline in learning represents a major challenge to education. The recent work on test-potentiated new learning, however, has shown that inserting brief tests into a lecture (or other extended study sequences) may serve as an antidote to this detriment and help students sustain a consistent level of learning throughout the lecture (Szpunar, Khan, & Schacter, 2013).

Perhaps due to its educational relevance, research interest in testing and its influence on new learning has risen in recent years. Figure 1 displays the number of publications included in

the current meta-analysis separated by publication year. A striking pattern can be seen. Although several studies were published during the end of the verbal learning era (e.g., Allen & Arbak, 1976; Tulving & Watkins, 1974), they were followed by a near complete absence of research on the topic until the late 2000s. However, a sustained and considerable interest in the phenomenon has developed over the past decade.<sup>1</sup>

A variety of experimental paradigms have demonstrated the beneficial effects of testing on new learning (see Figure 2). The most commonly used design is shown in Figure 2a. In this procedure, participants first encode a set of items, and then either perform retrieval practice on that set of items or not. All participants then encode another set of items, after which their memory for this second set of items is tested. For exposition purposes, we refer to the first learning episode as *original learning*, the retrieval practice phase for original learning as the *initial test*,<sup>2</sup> the second learning episode for a different set of items as *new learning*, and the memory test that assesses new learning as the *critical test*. In this design, original learning, initial testing, and new learning occur in separate trial blocks (Chan & McDermott, 2007; Szpunar et al., 2008). In contrast to this blocked design, researchers have sometimes implemented the initial test and new learning phases in a single trial block by interleaving retrieval practice trials with new learning trials (see Figure 2b). Here, participants first learn a set of items, after which they are asked to recall a particular studied item. They then encode a new item, then recall another previously studied item, then encode another new item, etc. (Davis &

---

<sup>1</sup> Note that Figure 1 shows only effect sizes that were included in the meta-analysis; so the total number of studies that have investigated the general phenomenon of test-potentiated new learning is higher than shown here (in the section on Inclusion/Exclusion Criteria, we detail why some studies were omitted from our analyses).

<sup>2</sup> Throughout this paper, as per the convention in the literature, we will use the terms “interpolated test,” “initial test,” and “retrieval practice” interchangeably.

Chan, 2015; Finn & Roediger, 2013).

In addition to the aforementioned multi-list design, researchers have also demonstrated the benefits of testing on new learning using a single-list design (Kornell & Vaughn, 2016). Here, participants only learn one set of items – i.e., there are no original learning trials – and participants are either tested before new learning or not. This type of design is commonly termed *pretesting*. Similar to the multi-list paradigm, the pretest and new learning trials can happen in separate blocks of trials (see Figure 2c) or in a single, interleaved block of trials (see Figure 2d). In this paradigm, participants are asked to guess the identity of the to-be-learned item when given a cue (e.g., whale - ?) before the target is shown for encoding (e.g., whale – mammal). Unlike studies in the generation effect procedure, in which participants are expected to generate the correct target, studies using the single-list pretesting procedure are designed to examine whether failing to retrieve an item can facilitate participants' ability to learn that item. Thus, materials are constructed to elicit an incorrect response during the pretest (by utilizing weak cue-target associations or obscure trivia questions). Because the item that participants retrieve (from semantic memory) during the pretest differs from the actual target, this paradigm allows one to examine the influence of prior retrieval on new learning in a single-list design.

When a multi-list learning design is used, both correct recall probability and number of intrusions can serve as dependent measures. Correct recall refers to the output of items that are studied during the new learning phase. Intrusions refer to the (erroneous) output of items that are studied during original learning when participants are instructed to recall items studied during new learning. When the single-list pretesting design is used, only correct recall can serve as the dependent measure.

A potential concern that confronts the present synthesis is whether it is sensible to

integrate results from several paradigms under a single meta-analysis. For example, one may wonder whether experiments that use the multi-list and single-list procedure (or experiments that use the blocked and interleaved procedure) are studying the same phenomenon. Here, we propose that all of the studies included in the present meta-analysis examine a similar, albeit broad, research question – that is, what are the effects of retrieval on subsequent learning of new information? Although different types of retrieval tasks may potentiate new learning based upon different component processes, researchers have also proposed that retrieval, no matter its nature (e.g., episodic or semantic), may affect new learning based on the same general mechanism (Divis & Benjamin, 2014; Finn, 2017; Kornell & Vaughn, 2016; Pastötter, Schicker, Niedernhuber, & Bauml, 2011).

As will become evident later, researchers have proposed myriad explanations to account for this phenomenon – however, many of these accounts were designed to handle study-specific findings. Although these specific explanations are valuable, they are typically narrow in scope, and we believe that future work on this phenomenon will benefit from the introduction of broader frameworks. An important objective of the present analysis is to organize existing explanations into broader theoretical frameworks. We then use the results of the meta-analysis to evaluate predictions emerging from these frameworks, thus providing, for the first time, a major comprehensive theoretical analysis of the existing data. Moreover, the diverse methodology involved in this sample of studies affords a crucial advantage for theory testing. For example, if a theory predicts that one should observe a larger benefit of retrieval on new learning in a multi-list procedure than in a single-list procedure, one can only test these predictions if the sample contains a sufficient number of studies for both procedures – a homogenous sample that lacks procedural diversity would make such theory testing very difficult. Nonetheless, in the present



paper, we take a dualistic approach by presenting both an omnibus meta-analysis that treats the literature as a unified whole for theory testing, and also present several subsample analyses that focus on more homogenous data sets.

### Historical Beginnings

Although the beneficial effects of prior retrieval on subsequent *relearning* have been widely known as "test-potentiated learning", there is no standard term yet for the benefits of prior retrieval on subsequent *new* learning. Indeed, in just a few years, researchers have used a variety of terms to describe essentially the same phenomenon. These terms include the *interim test effect* (Wissman, Rawson, & Pyc, 2011), the *forward testing effect* (Pastötter & Bauml, 2014), and *test-enhanced new learning* (Davis & Chan, 2015), and at other times researchers simply describe the finding without naming it (Szpunar et al., 2008). In the present paper, we use the term "test-potentiated new learning" because neither "interim test effect" nor "forward testing effect" specify the difference between relearning and new learning, and test-potentiated new learning is a natural extension of the well-known phenomenon of test-potentiated (re)learning of previously studied materials (Arnold & McDermott, 2013b; Izawa, 1971).

To our knowledge, the first reported occurrence of test-potentiated new learning (TPNL) can be traced to an unpublished study by Tulving, Patterson, and Malis (1967), which was cited in Malis' dissertation (1970). The effect's discovery occurred under somewhat serendipitous circumstances (Tulving, personal communication). Tulving and Malis had participants learn two sets of weakly related word pairs (i.e., an AB-AC learning paradigm) in a *single trial* for each set. To Tulving's surprise, participants recalled more words from original learning (i.e., the A-B list) than from new learning (i.e., the A-C list), which contradicted the typical results in the verbal learning literature (Allen & Arbak, 1976; Bruce & Murdock, 1968; Goggin, 1966). Malis

(1970), along with Tulving and Watkins (1974), eventually discovered that this pattern of results occurred because their initial procedure omitted the customary interpolated test between original and new learning, and that inclusion (or omission) of the interpolated test had a profound influence on new (i.e., A-C) learning.

### **Theoretical Perspectives**

To date, no formal theory has been proposed to account for the beneficial influence of testing on new learning, although researchers have provided a variety of explanations to account for individual findings. In a recent paper, Pastötter and Bauml (2014) provided an excellent overview of some of these accounts (see also Finn, 2017; Kornell & Vaughn, 2016); however, given the rapid development of research in this area, we believe that a more comprehensive examination is in order. To facilitate theoretical development on this topic, we have organized existing explanations into four categories that include resource theories, metacognitive theories, context theories, and integration theories. For exposition purposes, we refer to these different classes of explanations as "theories," although they are more similar to theoretical frameworks than formal theories. Some researchers have categorized existing accounts into encoding- or retrieval-based explanations (Cho et al., 2017; Pastötter & Bauml, 2014); we chose a different approach because many of the existing explanations do not fit neatly into encoding or retrieval processes, as differences in encoding often influence retrieval processes. We further note that these four classes of accounts *are not mutually exclusive*, even though their logic may at times offer different predictions about the influence of the same moderator variable. Indeed, given the broad and complex nature of the test-potentiated new learning phenomenon, it is possible, perhaps even likely, that multiple mechanisms underlie its occurrence. The primary purpose of the current theoretical review is to organize existing disparate explanations into major,

distinguishable categories of ideas, which is necessary to allow evaluation and comparison of these ideas within the framework of the current meta-analysis.

### **Resource theories.**

Explanations in this category generally claim that testing potentiates future learning by increasing the cognitive resources available for encoding new information, which can be achieved by *reducing* proactive interference (from original learning) during the encoding of new materials or by *increasing* the attentional resources that are available for post-retrieval encoding operations.

In a multi-list learning environment, items studied during original learning may intrude during new learning and thus interfere with one's ability to learn the new material. According to the logic of resource theories, interpolating a test between two encoding episodes should reduce the likelihood of intrusions from original learning during new learning, thus insulating the latter from encoding-based proactive interference (Darley & Murdock, 1971; Malis, 1970; Tulving & Watkins, 1974). The exact mechanism by which testing achieves this insulation effect awaits clarification, although several possibilities have been proposed. For example, the initial test may serve to reassure participants that they could remember the original-learning items (Malis, 1970). Alternatively, taking a test may provide the learner with cognitive closure (Roets, Van Hiel, & Cornelis, 2006; Webster & Kruglanski, 1994). Both of these possibilities may reduce the desire for learners to hold the original-learning items in working memory, thus availing more cognitive resources for the encoding of new materials. Consistent with this idea, recent work on expressive writing has shown that turning internal thoughts into external outputs can reduce subsequent intrusions from these thoughts (Klein & Boals, 2001; Park, Ramirez, & Beilock, 2014; Pennebaker & Chung, 2011; Ramirez & Beilock, 2011). Similarly, outputting items studied

during original learning in a memory test may reduce the likelihood that these materials will interfere with subsequent learning.

Other variants of resource theories suggest that testing may increase the cognitive resources available for future learning because it reduces mind wandering during the encoding task (Pastötter et al., 2011; Szpunar et al., 2013; Szpunar, Jing, & Schacter, 2014). Over the course of an encoding phase, participants may experience lapses of attention, which can have a profound impact on episodic encoding (Smallwood & Schooler, 2015). Interpolating a study phase with memory tests, however, may help sustain or redirect participants' attention to the encoding task, thus facilitating learning activities that occur later in an encoding sequence. Evidence for this idea comes from both electrophysiological and behavioral results. For example, Pastötter and colleagues (2011) measured brain oscillations in the alpha frequency band (8 - 14Hz) while participants learned five lists of words either with or without an initial test after each list. Their findings showed that alpha band activity, which is associated with memory load (Jensen, Gelfand, Kounios, & Lisman, 2002) and inattention (S. Palva & Palva, 2007), increased across the learning phase in the absence of interpolated testing. This rise in alpha power, however, was eliminated with interpolated testing. These findings led Pastötter and colleagues to suggest that testing reduced inattention by "resetting" the encoding process. In fact, Pastötter and colleagues found that a broad range of retrieval activities, including episodic recall, semantic generation, and 2-back (working memory) recognition, all produced similar electrophysiological (i.e., elimination of the rise in alpha power) and behavioral effects (i.e., test-potentiated new learning). Further evidence for resource theories has come from studies that used the random probe technique. For example, Szpunar and colleagues (2013) found that participants reported fewer bouts of mind wandering while watching a four-segment statistics video lecture if they

completed a test between each video segment. In contrast, Jing and colleagues (Jing, Szpunar, & Schacter, 2016b) showed that although interpolated testing failed to reduce the overall frequency of mind wandering when participants watched a lecture video on public health, it did qualitatively alter the type of thoughts in which participants engaged when they mind wandered. Specifically, participants who were tested intermittently reported more lecture-relevant thoughts than did nontested participants. In sum, interpolated testing may help learners sustain attention to the lecture content, which in turn enhances learning.

### **Metacognitive theories.**

Explanations in this category suggest that prior testing enhances later learning because it allows learners to optimize their encoding strategies, which is mediated by the metacognitive knowledge that one gains from having attempted retrieval of the original learning material. For example, taking a practice test may alert subjects to the type of information that is needed for successful performance during a criterial test (e.g., free recall requires one to self-generate retrieval cues, successfully recalling weakly related words may require one to encode the relational elements of those words), which allows subjects to better tailor their encoding during subsequent trials to suit those retrieval requirements (e.g., Chan, Manley, Davis, & Szpunar, 2018; Finley & Benjamin, 2012; Sahakyan, Delaney, & Kelley, 2004; Wissman et al., 2011).

Testing may also potentiate new learning by improving subjects' metacognitive awareness regarding their (in)ability to learn the target information. Indeed, without overt retrieval, learners are often overconfident (Dunlosky & Connor, 1997; Kang, 2010; Karpicke, 2009; Little & McDaniel, 2015). At the very least, an initial test can serve as a "reality check" for participants – to the extent that they are surprised by their underperformance, they may exert more effort during subsequent learning (Cho et al., 2017; H. S. Lee & Ahn, 2017). Testing can

also increase the effort that participants' expend on subsequent learning trials by altering their expectation about whether and when their memory will be tested. For example, in a multi-list learning environment, having taken a recent memory test increases learners' expectation that they will again be tested in the immediate future, even when they are told that whether a test will follow each study list is determined randomly (Weinstein et al., 2014). Such test expectancies have been shown to significantly influence how participants approach the encoding task (Balota & Neely, 1980; May & Thompson, 1989; Szpunar, McDermott, & Roediger, 2007).

Prior testing may also change how participants distribute their attentional resources during subsequent learning. For example, following an initial test, participants may devote more study time or effort (Dunlosky & Ariel, 2011; Son & Kornell, 2008) to items that they perceive as more difficult (Davis & Chan, 2015; Soderstrom & Bjork, 2014) or more important to learn (Chan, Manley, & Lang, 2017; Chan, Thomas, & Bulevich, 2009; Gordon & Thomas, 2014; Gordon, Thomas, & Bulevich, 2015; LaPaglia & Chan, 2013; LaPaglia, Wilford, Rivard, Chan, & Fisher, 2014; A. K. Thomas, Bulevich, & Chan, 2010; Wilford, Chan, & Tuhn, 2014), or they may use qualitatively different encoding processes, such as those emphasizing the semantic characteristics of the material (Chan et al., 2018; Soderstrom & Bjork, 2014). These changes in encoding processes are hypothesized to enhance subsequent learning of new material.

Metacognitive accounts can also be applied to test-potentiated new learning in the single-list procedure (see Figures 2c and 2d). For example, participants may gain insight into the type of conceptual relationship (i.e., the items are weakly related) present in cue-target pairs when they fail to generate the correct target. They may then employ encoding strategies that facilitate subsequent learning of these relations (e.g., by attempting to use imagery rather than simply using the preexisting, but weak conceptual relation between the words), a strategy change that is

unlikely to occur when participants simply study the weak associations.

On the surface, these metacognitive effects of testing on new learning appear to overlap somewhat with the resource account, so it is important to clarify their differences here. Resource accounts suggest that, without prior testing, continuous encoding causes a degradation in attentional resources (e.g., due to the buildup of proactive interference), and retrieval *restores* encoding resources to the optimal level. That is, when learners have performed retrieval before new learning, they are able to devote greater attention to the encoding task than if they had not performed retrieval. Metacognitive accounts, however, suggest that prior retrieval may enhance new learning by improving participants' encoding strategies, which may include better use of available attentional resources (e.g., by using deeper encoding or focusing on learning some items in lieu of others) – regardless of whether those resources had been depleted by prior learning. That is, unlike resource theories, which suggest that prior testing increases the *amount* of attentional resources available for encoding, metacognitive theories suggest that prior testing alters *how* learners make use of the available attentional resources (without affecting the amount of resources available). Another important distinction between these accounts is that resource theories ascribe test-potentiated new learning to the retrieval process itself. Specifically, retrieval automatically reduces mind wandering and restores attentional resources to the optimal level without requiring additional user input post-retrieval. In contrast, according to metacognitive theories, test-potentiated new learning occurs because learners use the metacognitive knowledge they gain from retrieval practice and apply it to new learning, which suggests a more conscious, effortful process.

### **Context theories.**

Explanations in this category suggest that testing enhances new learning by isolating the

original learning episode from the new learning episode (Chan & McDermott, 2007; Jang & Huber, 2008; Szpunar et al., 2008), which in turn enhances learners' ability to constrain their retrieval to specific memory sets. Context accounts have been applied extensively to explain a variety of memory phenomena. The core idea is that when people encode information, the content of that information (e.g., the meaning, the sound) is stored along with its study context. When one needs to retrieve that information later, the contextual information that is accessible to the learner can affect the likelihood with which one retrieves the target information (Godden & Baddeley, 1975; Lehman, Smith, & Karpicke, 2014; Morris, Bransford, & Franks, 1977; Tulving & Thompson, 1973; Sahakyan, Delaney, Foster, & Abushanab, 2013). Most important for present purposes, researchers have argued that attempting retrieval causes an internal context change relative to the study context (Abel & Bauml, 2016; Howard & Kahana, 2002; Jang & Huber, 2008; Jonker, Seli, & MacLeod, 2013; 2015; Pastötter et al., 2011). This context change can occur in two ways. First, attempting retrieval may induce one to enter a retrieval mode (Davis, Chan, & Wilford, 2017; Finn, 2017; Tulving, 1983), thereby establishing a retrieval context that differs from episodic encoding. Second, when items are recalled during retrieval practice, these items are updated with both the study context and the retrieval context (Chan, Erdman, & Davis, 2015; Jonker et al., 2013; Lehman et al., 2014; Whiffen & Karpicke, 2017). Unlike retrieval practice, it is generally hypothesized that restudying does not initiate a context change (because original study and restudy presumably involve the same, or very similar, encoding processes), so restudied items should not undergo context updating – that is, these items should be associated with only an encoding context, but not both an encoding and a retrieval context.

The logic of context change can be applied to explain TPNL in the multi-list design as



follows. Separating original- and new-learning with a test should help to isolate the former from the latter, because testing causes a context change. In the absence of retrieval practice, all studied items are associated with the study context. When participants are tested on the original-learning items before they are presented with new-learning items, the original-learning items are associated with both the study and retrieval contexts; in contrast, the new-learning items are associated with only the study context. This distinction in context association may facilitate subsequent retrieval of the new-learning items by allowing participants to constrain their memory search set to items that are associated with only the study context, therein reducing the pool of retrieval candidates (Brewer, Marsh, Meeks, Clark-Foos, & Hicks, 2010; Chan & McDermott, 2007; Chan, Wilford, & Hughes, 2012; Jacoby, Shimizu, Daniels, & Rhodes, 2005; Pierce, Gallo, & McCain, 2017; Shimizu & Jacoby, 2005; R. C. Thomas & McDaniel, 2013; Verhoeijen, Tabbers, & Verhage, 2011).

There are some key similarities between context theories and the aforementioned resource theories. From the perspectives of both theories, interpolated testing has the effect of isolating the original learning episode from the new learning episode, but they differ on how this isolation influences new learning. According to resource theories, isolation benefits new learning by increasing the attentional resources available to the learner during encoding. In contrast, context theories posit that the advantage of list isolation is revealed when participants attempt to retrieve the new-learning items during the criterial test.

### **Integration theories.**

Accounts in this category attribute test-potentiated new learning to enhanced integration of the new-learning material either with its retrieval cue or with the original-learning material. In one version of this theory, researchers suggest that testing increases the accessibility of original-

learning items; when participants attempt to learn new items that are related to the original items (e.g., if they are associated with each other by meaning, Chan & LaPaglia, 2013; or if they share a retrieval cue, Wahlheim, 2015), prior testing increases the likelihood that the original items would come to mind spontaneously during new learning. That is, prior testing increases the likelihood that *study phase retrieval* of the original-learning items would occur when participants study the new-learning items. This covert retrieval is hypothesized to promote binding of the original and new information into an integrated representation (Hintzman, 2004; 2009; Jacoby, Wahlheim, & Kelley, 2015; Nelson, Arnold, Gilmore, & McDermott, 2013; van Kesteren, Brown, & Wagner, 2016), which in turn facilitates later retrieval of the new-learning items by improving conceptual organization (Congleton & Rajaram, 2011; Jing, Szpunar, & Schacter, 2016a; Zaromb & Roediger, 2010) and the effectiveness of retrieval cues at the time of test (Carpenter, 2011; Pyc & Rawson, 2010). Other variants of integration theory exist. For example, researchers have suggested that attempting retrieval of previously learned materials may trigger a transient memory updating mechanism, which in turn facilitates the integration of new knowledge into the existing memory network (Chan et al., 2012; Chan & LaPaglia, 2013; Finn & Roediger, 2013; Hupbach, Gomez, & Nadel, 2009; 2011; J. L. C. Lee, Nader, & Schiller, 2017; Scully, Napper, & Hupbach, 2016; St Jacques, Olm, & Schacter, 2013).

Aside from promoting integration between the new-learning and original-learning materials, testing may also facilitate integration between the new-learning material and its retrieval cue. For example, when participants perform semantic retrieval in the single-list learning paradigm, having to guess the identity of a target before studying it (e.g., mammal - ?, Kornell, Hays, & Bjork, 2009) may pre-activate the target (mammal - whale) via spreading activation, thereby enhancing learning of the target (Grimaldi & Karpicke, 2012; Hays, Kornell,

& Bjork, 2013; Vaughn & Rawson, 2012). Even if the target is not directly activated during the semantic memory pretest, activation of the semantic network may prime it for memory updating and promote integration between the retrieval cue and the target. Attempting to answer a question may also cause participants to perform an elaborative memory search (Carpenter, 2011; Chan, McDermott, & Roediger, 2006; Pyc & Rawson, 2010), which produces long-term retention *of the question itself* along with the product of retrieval (Chan et al., 2006; Chan & LaPaglia, 2011; Kornell, 2014; Richland, Kornell, & Kao, 2009). When participants later encounter new information relevant to this question, learning is facilitated because the answer is better integrated with the question.

Recently, Finn (2017) proposed an episodic updating account, according to which retrieval can either potentiate or impair new learning depending on the learner's expectations. Following a retrieval trial, if learners deem the new information relevant to the present learning goals (e.g., if the new information is presented as feedback, as in the single-list paradigm, or if the new information is conceptually related to the recently retrieved information), the retrieval event is hypothesized to persist momentarily – this extended retrieval event is hypothesized to facilitate integration between the new and original information (or between the retrieval cue and the target). In contrast, if the new information is perceived to be irrelevant to the current learning goals or the recently retrieved information, the new information is hypothesized to not extend the retrieval event, and prior retrieval may in fact inhibit new learning (for a conceptually similar idea, see Chan & LaPaglia, 2013). Finn further argued that episodic retrieval should have a more powerful influence on new learning than semantic retrieval, because the former engages episodic retrieval mode (Davis et al., 2017), which, according to Finn, should facilitate the integration of new information into the episodic network.

## **Overview of the Meta-Analysis**

In the following, we present a quantitative summary of the literature on test-potentiated new learning. Our objectives are threefold: 1) to provide a comprehensive empirical summary of the available literature, 2) to examine variables that moderate the influence of retrieval on new learning and use these results to further theoretical development, and 3) to highlight areas with deficient empirical knowledge.

In the Method section, we first detail our literature search method and analysis strategies; we then describe coding approaches for all moderators. For organization purposes, we have classified the moderators as either theoretically-motivated or empirically-motivated. As is common for meta-analytic approaches, moderator analyses are designed to examine main effects and major patterns in the extant data – while they can speak to the nature of certain interactions, they cannot replace careful empirical investigations for the purpose of examining interactions between multiple variables, which remain instrumental for more precise hypothesis testing and theory development.

In the Results section, we begin by presenting the basic meta-analytic results. Afterwards, we evaluate the theories based on results of the moderator analyses using both a conventional, qualitative interpretation and then a quantitative approach guided by the results of a novel meta-regression. Briefly, the basic notion of this meta-regression is that studies with more factors that are predicted to promote TPNL (based upon a given theory) should produce a larger TPNL effect than studies that contained fewer such factors. This meta-regression approach has been used successfully to evaluate theories in the past (Michie, Abraham, Whittington, McAteer, & Gupta, 2009), and we have adopted their method here. More details regarding the logic of the meta-regression are presented in the Results section.

## Method

### Literature Search and Inclusion/Exclusion Criteria

One challenge in conducting the current review was that the phenomenon of test-potentiated new learning lacks an established, widely used term. Many studies have investigated the effects of prior retrieval on subsequent learning, but few have explicitly used the term “test-potentiated (new) learning” when describing the phenomenon. As a result, we used a variety of terms in our literature search. The search engine used was PsycINFO.<sup>3</sup> The terms (searched under “Anywhere”) and the number of results generated as of April 15, 2016 are displayed in Table 1. The first and second authors examined all of the results, then further located relevant articles based on the reference sections of the included studies. In addition, we emailed research groups that have published in this domain to inquire about unpublished studies. Twenty-eight of the 29 contacted research groups responded to our query and 17 provided unpublished data.

Studies were selected for inclusion in the meta-analysis based on the following inclusion criteria. First, studies must have assessed the influence of retrieval on learning *new* information. Specifically, in at least one experimental condition, subjects would perform some form of retrieval practice (e.g., episodic retrieval or semantic retrieval) before they learned new information that *differed from what they had previously attempted to retrieve*. Second, studies must have measured memory performance for the new-learning material following retrieval relative to a control condition that did not include retrieval. Finally, studies must have provided sufficient statistical information to compute the TPNL effect.

Exclusion of studies from the meta-analysis generally fell into the following categories:

---

<sup>3</sup> We replicated many of these searches using Google Scholar; however, because Google Scholar lacks the filtering options of PsycINFO, the number of results it produced were often far too large (e.g., in the thousands) for systematic reviews.

1) The study investigated test-potentiated *relearning* instead of test-potentiated *new* learning (e.g., Arnold & McDermott, 2013a; Izawa, 1967; 1970; 1971; Karpicke, 2009; Little & McDaniel, 2015; Nelson et al., 2013; Young, 1971; Yue, Soderstrom, & Bjork, 2015). These studies were excluded because they conflated the testing effect with test-potentiated learning, making it difficult to disentangle the benefits conferred by retrieval practice itself (i.e., direct effect of testing) and those produced by test-potential of the relearning episode (i.e., indirect effect of testing). 2) The study investigated the influence of retrieval practice on *new learning*, but there was no straightforward way to assess the potentiating effects of retrieval free from other confounding factors (Brewer et al., 2010; Chan & McDermott, 2007; Malis, 1970; Mayer et al., 2009; Sahakyan et al., 2004; Verkoeijen et al., 2011; Weinstein, Nunes, & Karpicke, 2016). For example, in Chan and McDermott (2007) and Brewer et al. (2010), participants studied two different word lists and were either tested initially on both lists or not. Later, all participants received a final test for both lists. Similar to the test-potentiated relearning studies above, this methodology conflated testing effect with test-potentiated new learning, because the tested participants might outperform the control participants on List 2 (i.e., new learning) because they were initially tested on List 1 (i.e., TPNL) or because they were initially tested on List 2 (i.e., testing effect). 3) The study featured a learn-to-criterion method for materials studied after the initial test, which masked the effects of test-potentiated new learning (Hupbach et al., 2009; e.g., Hupbach, Gomez, Hardt, & Nadel, 2007; Potts & Shanks, 2012). 4) The study met the inclusion criteria, but insufficient data were available to compute an effect size (Wissman & Rawson, 2015b Experiment 7).

The final set of studies included a total of 159 independent effect sizes, of which 126 were described across 42 publications, and 33 were acquired from unpublished dissertations or

directly from authors. The entire set of studies included data from 8,767 participants. Table S1 in the supplementary materials presents the sample size and effect size for each independent sample.

### **Coding of Theoretically-Motivated Moderators**

#### **Interleaving retrieval practice with new learning.**

A primary moderator of interest is whether the initial test and new learning trials were presented in a blocked (see Figures 2a and 2c) or an interleaved fashion (see Figures 2b and 2d). Several researchers (Bettencourt, Delaney, & Chang, 2015; Davis et al., 2017; Davis & Chan, 2015; Finn & Roediger, 2013) have shown that intermixing retrieval practice trials with new learning trials can reverse the typical beneficial influence of testing on new learning. Accordingly, studies were coded for whether an *interleaved* or a *blocked* design was employed.

#### **Research design (within-subjects or between-subjects).**

The use of a within-subjects vs. between-subjects design has been shown to influence a variety of memory phenomena (e.g., for the generation effect, Bertsch, Pesta, Wiscott, & McDaniel, 2007; for the distinctiveness effect, Huff, Bodner, & Fawcett, 2015; for the production effect, MacLeod, Gopie, Hourihan, Neary, & Ozubko, 2010; for the testing effect, Rowland, 2014). We therefore investigated the impact of this variable on test-potentiated new learning. Studies that randomly assigned participants to either retrieval practice and control conditions were coded as *between-subjects*, and the remaining studies were coded as *within-subjects*.

#### **Comparison task.**

Here we examine whether the beneficial effects of testing on future learning vary by comparison tasks (i.e., the non-retrieval task employed in the control condition). We divided

comparison tasks into four categories: restudy, filler, no-test, and math problems. In the *restudy* category, participants in the control condition either experienced a re-presentation (e.g., Davis & Chan, 2015; Finn & Roediger, 2013) or an extended presentation (Potts & Shanks, 2014) of the materials. In the *filler* category, the comparison task engaged participants on learning-irrelevant material, such as drawing pictures (Allen & Arbak, 1976; Arkes & Lyons, 1979; Tulving & Watkins, 1974), playing a video game (Chan et al., 2009; Gordon & Thomas, 2014), searching for specific words/numbers (Lane, Mather, Villa, & Morita, 2001; Robbins & Irvin, 1976), etc. In the *no-test* category, participants did not perform a distractor task, nor did they restudy the to-be-remembered materials (Robbins & Irvin, 1976; Tulving & Watkins, 1974; Wahlheim & Jacoby, 2010; Wissman et al., 2011). In the *math problems* category, participants were required to perform mathematical operations such as mental algebra or counting backwards by 3 (Divis & Benjamin, 2014; Nunes & Weinstein, 2012). This math category was included because of its popularity as a filler task and the high degree of cognitive load it may place on the learner.

### **Relation between original and new learning.**

Across the present sample of studies, some showed related single words across lists (e.g., "squirrel" in List 1 and "zebra" in List 2, Szpunar et al., 2008), some used weakly related pairs with no direct association between the original-learning and new-learning targets (e.g., pearl-harbor in List 1 and pearl-jewelry in List 2, Wahlheim, 2015), some used unrelated paired associates across lists (e.g., pearl-harbor in List 1 and pearl-jewelry in List 2, Wahlheim, 2015), and some used unrelated single words across lists (Pastötter et al., 2011). We coded relatedness based on whether there were pre-existing associations between the target items across lists. Using this criterion, the first example was classified as *related* while the last three were classified as *unrelated*.



### **Initial test format.**

Initial test format refers to the type of test that participants take before new learning. This variable was coded across five levels. *Item cued recall* referred to memory tests in which participants were cued to recall a *specific piece* of information, such as when participants were asked to recall the target of a particular paired-associate (Allen & Arbak, 1976; Davis & Chan, 2015; Picho, Rodriguez, & Finnie, 2013; Wahlheim, 2015). *Free recall* refers to a test in which participants were not provided with any retrieval cues, and it often occurred when participants learned only two sets of items, and the initial test occurred following encoding of the first set of items (Lane et al., 2001; Wissman & Rawson, 2015a; 2015b). *List cued recall* was similar to free recall, except that participants were required to recall a *specific set* of items. This task was frequently used when participants studied more than two sets of items, and were told to recall the most recent set during an initial test (Bauml & Kliegl, 2013; Pashler, Kang, & Mozer, 2013; Szpunar et al., 2008). *Nonepisodic recall* required participants to retrieve items not relevant to the to-be-learned materials, including semantic generation (Divis & Benjamin, 2014), N-back working memory (Pastötter et al., 2011), and autobiographical recall (Pastötter & Bauml, 2016; Weinstein, McDermott, Szpunar, Bauml, & Pastötter, 2015). *Pretest* referred to a type of nonepisodic retrieval whereby subjects guessed the identity of a target before they studied it – i.e., these studies used the single-list procedure shown in Figures 2c and 2d. Following the guess (or semantic retrieval), participants studied the target as the new learning material.

### **Test format match**

We coded whether the test format for the initial and criterial tests *matched or not*. Note that studies that used a nonepisodic retrieval task (e.g., semantic generation interpolated between original- and new-learning, a pretest that occurs before new-learning) during the initial test were

coded as non-matched even if the same test was used during both the initial and criterial tests (e.g., item cued recall, whale - ?), because here the initial test requires retrieval from semantic memory, whereas the criterial test requires retrieval from episodic memory.

**Memory load (sets of materials studied before new learning).**

We included the number of lists (or sets) of materials studied as a moderator. For exposition purposes, we refer to this moderator as memory load. Across the 159 effect sizes in the sample, 45 used a single-list (pretesting) design, 60 used a two-list design, 10 used a three-list design, 21 used a four-list design, 21 used a five-list design, and two used a 12-list design. We opted to use number of lists as the moderator, rather than number of items within list or total number of items across lists for two reasons. First, from a theoretical perspective, initial testing separates each *set* of study items (rather than within a set of items), so its influence should be particularly evident across item sets rather than across individual items. Second, although it is relatively straightforward to obtain a total number of items or within-list number of items for word lists, the data that are required to obtain such a word count are unavailable for more complex materials such as prose or videos.

**Retrieval practice performance.**

We examined whether retrieval practice performance moderated the magnitude of test-potentiated new learning. For this moderator analysis, we excluded studies that did not use episodic retrieval as the initial test (Divis & Benjamin, 2014; Kornell et al., 2009; Pastötter et al., 2011; Potts & Shanks, 2014) because performance on nonepisodic retrieval tasks do not provide an index of original learning. We also excluded studies that did not report performance in proportion units or for which proportion recalled could not be estimated.

**Delay between original learning and new learning.**

This delay refers to the time interval (in seconds) between the end of the original learning phase and the start of the new learning phase amongst studies that used the multi-list design. For those that used the single-list design, it refers to the time between the pretest of a given item and encoding of the target item. For experiments in which participants studied more than two lists (Szpunar et al., 2013; Weinstein, McDermott, & Szpunar, 2011), we coded the interval between the last two studied lists. This delay varied considerably among the included studies, ranging from none (Grimaldi & Karpicke, 2012; Kornell et al., 2009; Wissman & Rawson, 2015b) to one week (Chan & LaPaglia, 2011). While the majority of studies used a relatively short delay ( $< 37$  min,  $k = 153$ ), and only six studies used a delay of at least 24 h. Given the large range of delay used, a logarithmic transformation was performed on this variable. For studies with a 0 s delay, which cannot be logarithmically transformed, we substituted “undefined” with 0.

**Coding of Empirically-Motivated Moderators****Publication status.**

In addition to computing fail-safe  $N$ , one way to examine possible publication bias is to compare the effect sizes of published and unpublished studies. In the current meta-analysis, we assessed whether effect sizes for published studies were greater than those for unpublished studies, a pattern that could indicate a bias against publishing nonsignificant effect sizes. We opted not to use a funnel plot as an interpretation tool for publication bias given its subjective nature (Terrin, Schmid, & Lau, 2005). Although the Egger test and the Begg test are sometimes used to quantify funnel plot data, their conclusions do not always align and thus can add to interpretive confusions (Lau, 2006). Given the number of unpublished data sets included in our sample, it was determined that using publication status as a moderator provides a clear and

objective way to detect the potential existence of publication bias.

### **Participant sample.**

Given increasing popularity of internet-based data collection tools such as Amazon Mechanical Turk (Buhrmester, Kwang, & Gosling, 2011; Paolacci, 2010; Topp & Pawloski, 2002), we examined whether effect sizes differed based upon whether online or laboratory-based samples were collected. Although studies in psychological science have generally reported similar findings regardless of whether data collection occurred online or in lab (Finn & Roediger, 2013; Simons et al., 2014), important differences have also emerged (Goodman, Cryder, & Cheema, 2013).

### **Material type.**

A range of materials have been used to study the influence of retrieval on new learning, including lists of single words (Pastötter et al., 2011; Szpunar et al., 2008), word pairs (Kornell et al., 2009; Potts & Shanks, 2014), picture-word pairs (Davis & Chan, 2015; Weinstein et al., 2011), trivia facts<sup>4</sup> (Kornell, 2014; Pashler et al., 2013), prose passages (Wissman et al., 2011), video materials (Chan et al., 2012; Gordon, 2016; Gordon & Thomas, 2014; Szpunar et al., 2013), and slides (Huff, Davis, & Meade, 2013). Given the variety of materials involved, we coded studies into the categories of *Words, Paired Associates* (which included word pairs and picture-word pairs), and *Prose/Videos* (which included facts, trivia questions, passages, slides, and videos).

---

<sup>4</sup> In all of the studies that used trivia facts as materials, the facts were obscure and meant to generate floor-level performance before participants encode the answer. Consequently, learning the correct answers to these obscure facts requires episodic encoding.

**Criterion test format.**

Criterion test format was coded into five categories, including list cued recall, free recall, item cued recall, modified-modified free recall (MMFR)<sup>5</sup>, and recognition. The first three test formats were defined in the manner described in the initial test format section. In *MMFR*, participants are given a cue and asked to recall as many targets associated with the cue as possible. This procedure was primarily used among studies in which participants had learned multiple targets with the same cue (Arkes & Lyons, 1979; Chan et al., 2009; Chan & LaPaglia, 2011; Gordon & Thomas, 2014; Pashler et al., 2013; Robbins & Irvin, 1976). *Recognition* included multiple-choice recognition (Divis & Benjamin, 2014; Potts & Shanks, 2014) and source recognition tests (Chan et al., 2012; Huff et al., 2013; Lane et al., 2001).

Some studies have issued multiple tests for the new-learning materials. For example, in Szpunar et al. (2008), participants were tested on the new-learning items 1 min after they studied them. Following a 30 min delay, participants were again given 8 min to recall all studied items, including both the original-learning items and the new-learning items. In such cases, we always treated the first memory test for new-learning items as the criterion test.

**Administration of corrective feedback.**

We coded this moderator based upon whether the correct answer was presented following retrieval during the initial test. Research on the testing effect has shown that feedback can greatly

---

<sup>5</sup> “Modified modified free recall,” which has taken on the unfortunate reputation of a poor way to name a task, is often mistakenly attributed to Barnes and Underwood (1959). Although Barnes and Underwood introduced the procedure, they were not responsible for its name. Underwood (1948), however, did establish the procedure and name for modified free recall (MFR), even though it is often (e.g., Wikipedia) incorrectly attributed to Briggs (1954). The first documented use of the MMFR term should be attributed to Melton (Tulving, personal communication), who wrote a “Comments” piece for a book chapter by Postman (1961). Here, Melton wrote, “it seems obvious that we need to exploit the ingenious free-recall techniques previously referred to as MFR and MMFR.” (Melton, 1961, p. 183)

increase the benefits of retrieval practice, so it is important to examine whether feedback also influences the magnitude of test-potentiated new learning. Study that provided corrective feedback to participants during retrieval practice were assigned to the *feedback* category.

### **Retention interval.**

Retention interval refers to the delay (in seconds) between the end of the new learning phase and the start of the criterial test. Among the included studies, the retention interval ranged from immediate to 7 days. A logarithmic transformation of this duration (in seconds) was conducted, with undefined (0 sec) intervals coded as 0.

### **Effect Size Calculations**

The primary outcomes of interest included correct recall probability and the number of intrusions, with the latter available only for studies that used the multi-list design. *Correct recall probability* refers to the proportion of studied items from new learning that are recalled during the criterial test. *Intrusions* refer to items from original learning that are erroneously recalled when participants are instructed to output only items from new learning.

Several studies provided means but not measures of variability. For these studies, when possible, standard deviations were computed using the reported *t*-value. For cases in which such estimations were not possible (e.g., when *t*-value was reported as “<1” or when *t*-values were not reported), we imputed variability based on the average *SD* from the available studies and marked them with an asterisk in the effect size column in the Appendix. We calculated separate average *SDs* for studies that manipulated testing between-subjects and those that manipulated testing within-subjects.

For correct recall data, nine effect sizes did not include any variability data (Arkes & Lyons, 1979; Robbins & Irvin, 1976; Tulving & Watkins, 1974). Among these studies, all but

one (Tulving & Watkins, 1974, Experiment 2) used a between-subjects design. The between-subjects *SD* from experiments that reported proportions data ( $k = 79$ ) was .19 for the tested condition and .20 for the control condition, which did not differ significantly,  $t(156) = .98$ ,  $p = .33$ . The equivalent within-subjects *SD* ( $k = 53$ ) was .20 for the tested condition and .21 for the control condition, which again did not differ,  $t(52) = 1.11$ ,  $p = .27$ . We assigned these *SDs* to the aforementioned effect sizes.

For intrusion data, we imputed *SD* for one study (Bauml & Kliegl, 2013), which manipulated testing within-subjects. We calculated the *SD* from the within-subjects experiments that reported these data in proportions ( $k = 3$ ). The average *SD* was .12 for both the tested and control conditions. We therefore assigned this value to the Bauml and Kliegl study. In addition, when subjects in the initial test condition produced no intrusions, the associated *SD* was 0. For these cases, we replaced 0 with 0.1 for effect size calculations (Wissman et al., 2011; Wissman & Rawson, 2015b).

Analyses were conducted using the software package Comprehensive Meta-Analysis (version 3). All effect sizes were calculated based on standardized differences in means (Cohen's  $d$ ). Because Cohen's  $d$  is a slightly biased estimate of true effect sizes (especially for small or skewed samples), we converted Cohen's  $d$  to Hedge's  $g$  using the following formula:

$$g = \left(1 - \frac{3}{4(df) - 1}\right)d$$

One complication is that studies varied in whether they used a within-subjects or between-subjects design. When calculating Cohen's  $d$  for within-subjects comparisons, it is necessary to adjust for the correlation between the tested and control condition scores. However, none of our included studies provided this statistic. To address this issue, we assumed a correlation of .50 for all within-subjects studies. The advantage of this assumption is that it

allowed the same Cohen's  $d$  formula to be used for both between-subjects and within-subjects designs - a solution recommended by Cumming (2012).

The included studies varied with respect to how many sets of materials participants had to study. Although a majority of the studies had participants learn two sets of materials, some had participants learn three or more sets. When this occurred, we used performance on the last set of materials to estimate effect size (but we also coded the number of material sets studied as a separate moderator). For example, in Experiment 2 of Szpunar et al. (2008), participants studied five lists of words. Participants in the *always-tested* condition received an immediate test following each list; however, the remaining participants received an immediate test only after List 2, List 3, List 4, or List 5 (which was manipulated between subjects). In this case, we computed the effect size by comparing immediate recall of List 5 between the *Always-Tested* condition and the *Tested-after-List-5* condition.

For all moderator analyses, we assumed a common among-study variance across subgroups (i.e., a random effects model). That is, we pooled the within-group estimates of tau-squared (i.e., the variance of effect sizes across studies within a level of the moderator variable) and applied this common estimate across all studies in a moderator analysis. We chose this approach because most studies used college students as participants, and because pooled estimates of tau-squared is the preferred method for a moderator analysis when some subgroups contain a small number of effect sizes (Borenstein, Hedges, Higgins, & Rothstein, 2009).

## Results

We first report descriptive statistics regarding the sample of studies under consideration before turning to results of the moderator analysis. To provide a comprehensive summary of the data on test-potentiated new learning, we examined the data in four ways. First, we assessed the



correct recall results from the complete sample of studies ( $k = 159$ , see Table 4). When appropriate, we also discuss the results from three subsample analyses. In the “standard procedure” subsample analysis, we consider the correct recall results from studies that used the most common procedure, in which participants completed episodic retrieval practice in the multi-list, blocked design ( $k = 84$ , see Figure 2a for the procedure and Table 5 for the results). In the intrusion subsample analysis, we consider the results for studies that reported intrusion data (i.e., a different dependent variable than correct recall,  $k = 41$ , see Table 6). In the pretesting subsample analysis, we report the correct recall results from studies that implemented the single-list procedure ( $k = 45$ , see Figures 2c and 2d for the procedure and Table 7 for the results). The point estimate of effect sizes ( $g$  for categorical moderators and  $B$  for continuous moderators), confidence intervals,  $Q$ -values (which indicates between-studies heterogeneity), and  $p$ -values are shown in Tables 4-7. For the sake of brevity, we report only point estimates in the text, except for follow-up analyses, which are not included in the Tables. All effect sizes are weighted and based upon a random effects model, and continuous moderators were examined with meta-regression using the method-of-moments random effects model.

### **Overall Effect Sizes and Influence of Moderators**

The *complete sample* showed a robust test-potentiated new learning effect on correct recall,  $g = 0.44$  (see Figure 3). Fail-safe  $N$  (Rosenthal, 1979) was 21,140. A stem-and-leaf plot is presented in Table 2, which shows a moderately negatively skewed (-0.65) unimodal distribution, although a majority of the effect sizes (79%) were positive. There was substantial variability in the data,  $Q = 1130.59$ , and between-studies variability contributed to most of the heterogeneity,  $I^2 = 86\%$ . The *standard procedure subsample* ( $k = 84$ ), which contained studies that used the multi-list, blocked design (see Figure 2a and the darker shaded cells in Table 2),

and in which the initial test involved an episodic retrieval task (i.e., not semantic retrieval or autobiographical retrieval of information unrelated to original-learning), yielded a robust test-potentiated new learning effect,  $g = 0.75$ , fail-safe  $N = 5,673$ , skew = 0.01,  $Q = 280.93$ ,  $I^2 = 70\%$ . Data from the *intrusion subsample* ( $k = 41$ ) showed that retrieval substantially reduces intrusions,  $g = -0.77$ , fail-safe  $N = 3,585$ , skew = -1.19,  $Q = 162.97$ ,  $I^2 = 75\%$ . Table 3 shows a stem-and-leaf plot of the data. This subsample included only studies that used the multi-list design, which were the only studies that could report intrusions. There is substantial overlap between the studies in this subsample and those in the standard procedure subsample (37 out of 41 studies in the intrusions subsamples also use the standard procedure). For the *single-list, pretesting subsample* ( $k = 45$ , see the light-shaded cells in Table 2), we observed a smaller test-potentiated new learning effect,  $g = 0.35$ , Fail-safe  $N = 1,658$ , skew = -0.37,  $Q = 361.29$ ,  $I^2 = 88\%$ . To provide a comprehensive overview of the data set, individual effect sizes for all studies are shown in the Appendix. While we discuss results of the most important moderators below, all moderator analyses are provided in Tables 4-7. The theoretical relevance of each moderator is displayed in Tables 8-11. In these Tables, we highlighted the experimental conditions that are expected to produce greater effect sizes based on the logic of each theoretical framework.

### **Results for Theoretically-Motivated Moderators**

#### **Interleaving retrieval practice and new learning.**

As can be seen in Tables 8-11, interleaving has theoretical relevance for all four theories. When initial testing and new learning were performed in separate blocks of trials, testing facilitated new learning,  $g = 0.67$ . Strikingly, when initial testing and new learning trials were intermixed, testing no longer potentiated new learning,  $g = -0.02$ . This finding is consistent with the hypothesis proposed by Davis and Chan (2015) and Finn (2017), who argued that requiring

learners to repeatedly switch between retrieval and encoding operations can incur a cost, instead of a benefit, on new learning. As will become obvious, interleaving has a profound impact on the magnitude of the test-potentiated new learning effect. In subsequent moderator analyses, when a particular level of the moderator was confounded with interleaving (e.g., most of the studies that used the interleaving procedure also used paired associates as their study material), we highlight the confound with an asterisk in Table 5. In such cases, an examination of the standard procedure and intrusion subsample data (i.e., Tables 6 and 7) can shed light on test-potentiated new learning absent the influence of interleaving, as these subsamples did not include studies that used the interleaving design.

In the single-list, pretesting subsample, presenting retrieval practice trials and new learning trials in an interleaved fashion had minimal influence on new learning,  $g = 0.37$ , relative to presenting these trials in separate blocks,  $g = 0.29$ ,  $Q = 0.23$ ,  $p = .63$ . Given the powerful influence of interleaving in the complete sample, this null effect may seem surprising. However, this pattern was accurately predicted by both metacognitive theories (Table 9) and integration theories (Table 11), whereas resource theories and context theories received only partial support from the data.

### **Research design.**

Research design has theoretical relevance for metacognitive theories (see Table 9 for the prediction), and the results of this moderator analysis supported the prediction. Specifically, studies that employed a between-subjects design showed a larger test-potentiated new learning effect,  $g = 0.56$ , than those with a within-subject design,  $g = 0.25$  (see Table 5), and the same pattern was observed in all of the subsamples (see Tables 6 - 8).

### **Comparison task.**

The moderator of comparison task has important implications for context theories (Table 10) and integration theories (Table 11). Initial testing enhanced new learning compared to doing math,  $g = 0.77$ , other filler tasks,  $g = .79$ , and no-testing,  $g = 0.64$ , but not when compared to restudying,  $g = 0.09$ . A majority of the experiments that used restudy as the comparison task, however, also interleaved testing with new learning (44 out of 66 samples). When these studies were removed from the analysis, as in the standard procedure subsample, the moderating effect of comparison task was no longer significant,  $Q = 2.32$ ,  $p = .51$ , and retrieval potentiated new learning when compared to restudying,  $g = 0.61$ . Similar results were also observed in the intrusion (Table 7) and pretesting subsamples (Table 8). We deemed these results as inconsistent with context theories (i.e., testing potentiated new learning relative to filler tasks) and partial support for integration theories (i.e., the TPNL effect was smaller when compared to restudy, at least in the complete sample).

### **Relation between original and new learning.**

This moderator has theoretical relevance for resource theories (Table 8), context theories (Table 10), and integration theories (Table 11). Testing conferred a greater benefit on new learning when the original- and new-learning materials were related,  $g = 0.61$ , than when they were unrelated,  $g = 0.25$ .<sup>6</sup> This pattern was also evident across all of the subsamples, although

---

<sup>6</sup> One may wonder whether the effects of relatedness would differ if we had included studies into the “related” category when the original-learning and new-learning items shared a retrieval cue, even if the targets were unrelated (e.g., pearl-harbor, pearl-jewelry, or a face-name pair and face-profession pair). To examine this possibility, we conducted an additional analysis using this alternative coding scheme. The results of this moderator analysis differed from the original one. Although relatedness still has a significant influence on the magnitude of test-potentiated new learning, the effect size was actually greater among the “unrelated” studies ( $g = 0.54$ ,  $k = 39$ ) than the “related” studies ( $g = 0.41$ ,  $k = 120$ ). Despite this result, we

the effects in the subsamples were smaller and did not reach significance (see Table 6-8). These results are largely consistent with the predictions from resource theories and integration theories, but not from context theories.

### **Initial test format.**

Initial test format has theoretical implications for resource theories (Table 8), metacognitive theories (Table 9), and integration theories (Table 11). All initial test formats facilitated new learning, but nonepisodic recall did not,  $g = 0.32$ ,  $p = .12$ . This nonsignificant effect may be attributed to the small number of studies in this level of the moderator ( $k = 8$ ). Further empirical investigations would be useful to determine whether, or to what extent, nonepisodic retrieval can enhance new learning. It is also worth noting the surprisingly small potentiating effect of item cued recall ( $g = 0.17$ ). Here again, many of the item cued recall studies also used the interleaving procedure ( $k = 22$  out of 54). When the data were examined from the standard procedure subsample, which did not include studies that employed the interleaving design, initial item cued recall potentiated new learning ( $g = 0.71$ ) to a similar degree as the other recall formats,  $Q = 0.87$ ,  $p = .65$  (see also data from the intrusion sample, Table 7). All three relevant theoretical frameworks accurately predicted that nonepisodic recall would not enhance new learning, but unlike resource theories, metacognitive theories and integration theories additionally predicted correctly that pretesting would benefit new learning.

### **Test format match.**

Test format match provides a test for metacognitive theories. Our data showed that the impact of this variable was marginal,  $Q = 3.21$ ,  $p = .07$ . Surprisingly, testing enhanced new

---

stayed with the original classification of this moderator because the concept of “shared cues” can be very broad, given that all items studied in a given experiment technically share the same temporal and environmental cues.

learning to a greater degree when the test formats mismatched ( $g = 0.50$ ) than when they matched ( $g = 0.33$ ), and this pattern extended to the standard procedure subsample ( $g_{mismatch} = 0.86$ ,  $g_{match} = 0.66$ ). Notably, the intrusion data demonstrated the opposite pattern ( $g_{mismatch} = -0.45$ ,  $g_{match} = -0.87$ ), thus revealing a somewhat rare dissociation between data for correct recall and intrusions. Altogether, these results are largely inconsistent with the prediction emerging from metacognitive theories.

### **Memory load.**

Memory load is relevant to resource theories (see Table 8). There was a small, but significant, positive association between memory load and the size of the test-potentiated new learning effect,  $B = 0.08$ ,  $p < .01$ . Namely, the beneficial effects of testing on new learning increased with the amount of materials participants studied prior to new learning (see Figure 4a). The range of this moderator variable was somewhat restricted—157 of the 159 samples involved five or fewer sets of items, and only two samples used a procedure in which participants studied 12 sets of items (Pashler et al., 2013). These latter studies might have had a disproportionate influence on the regression results (i.e., they were outliers in this moderator). However, removing these two studies from the meta-regression actually *increased* the effects of memory load,  $B = 0.15$ ,  $CI [0.08, 0.22]$ ,  $p < .01$ . Analysis of the standard procedure and intrusion subsamples suggested that memory load did not affect the magnitude of test-potentiated new learning,  $ps > .72$ ; however, when the aforementioned outlying studies (Pashler et al., 2013) were removed, marginal effects of memory load emerged for both the standard procedure subsample,  $B = 0.07$ ,  $CI [-0.01, 0.16]$ ,  $Z = 1.79$ ,  $p = .07$ , and the intrusion subsample,  $B = -0.12$ ,  $CI [-0.25, 0.02]$ ,  $Z = -1.67$ ,  $p = .10$ . Overall, these data are consistent with resource theories.

### **Retrieval practice performance.**

Both metacognitive theories (Table 9) and integration theories (Table 11) make predictions based on the moderator of retrieval practice performance. There was a positive association between initial test performance and the magnitude of test-potentiated new learning,  $B = 1.17, p < .01$  ( $k = 78$ ). It is important to note, however, that nearly half of the studies with low initial test performance (i.e.,  $M < .50$ ) also use the interleaved design (18 of 38 samples), which tended to produce a test-impaired new learning effect. Therefore, the positive association between retrieval practice performance and TPNL might be driven by this cluster of studies (see Figure 5, in which the interleaved studies are shown as filled circles). A different pattern emerged when we considered the results from the standard procedure subsample, which did not include the studies that used the interleaving procedure. Here, initial test performance was not associated with the magnitude of test-potentiated new learning,  $B = 0.15, p = .68$  ( $k = 56$ ). Similarly, the intrusion subsample data also showed that retrieval practice performance was not significantly associated with TPNL,  $B = .65, p = .17$  ( $k = 31$ ). The data from this moderator, therefore, were inconsistent with the prediction based upon both metacognitive theories and integration theories.

### **Delay between original learning and new learning.**

Delay between encoding episodes is relevant to resource theories (Table 8) and context theories (Table 10), and this moderator did not affect the magnitude of test-potentiated new learning (see Figure 6a),  $B = -0.04, p = .22$ . Because only six samples used a delay that was longer than 14 min (i.e., 24 hr in Pashler et al., 2013, and one week in Chan & LaPaglia, 2011), we also examined the impact of short-term delays on test-potentiated new learning by excluding these samples, and the results remained similar,  $B = -0.06, CI [-0.15, 0.03], Z = -1.28, p = .20$ .

Moreover, when we examined the data from the standard procedure and intrusion subsamples, a similar pattern emerged. Specifically, when the outlying influence of the very-long delay studies were removed from this analysis, delay did not affect the magnitude of test-potentiated new learning in either the standard procedure subsample (see Figure 6b),  $B = -0.06$ ,  $CI [-0.16, 0.05]$ ,  $Z = -1.02$ ,  $p = .31$ , or the intrusion subsample (see Figure 6c),  $B = 0.21$ ,  $CI [-0.28, 0.70]$ ,  $Z = 0.84$ ,  $p = .40$ . A different outcome, however, was found in the pretesting subsample. Here, delay was *negatively* associated with test-potentiated new learning, (see Figure 6d),  $B = -0.16$ ,  $p = .01$ . However, only one study (Kornell, 2014) implemented a delay that was longer than 10 min (i.e., 24 hr). The significant negative association between delay and the pretesting effect remained after removing this study,  $B = -0.27$ . Examination of Figure 6d suggests that the TPNL effect in the single-list procedure essentially disappears following a delay of less than 10 min. Nevertheless, it is noteworthy that Kornell showed a powerful TPNL effect after a 24 hr delay (see the rightmost data point in Figure 6d). In this experiment, Kornell used complex trivia questions (instead of paired associates) as the study material. Therefore, the relation between delay and TPNL in this paradigm may depend on material type. Taken together, the data from this moderator failed to support the prediction from both resource theories and context theories.

### **Results for Empirically-Motivated Moderators**

#### **Publication status.**

Published studies showed a larger test-potentiated new learning effect,  $g = 0.50$ , than unpublished studies,  $g = 0.22$ . At first blush, this finding indicates that smaller effects might be less publishable than larger effects; however, a mitigating factor appears upon closer inspection of the data. Specifically, among the 33 unpublished studies, 14 used the interleaving design. When these studies are removed from consideration, as in the standard procedure subsample,



publication status was no longer associated with different effect sizes,  $p = .93$ . Intriguingly, studies in the intrusion and pretesting subsamples also showed smaller effect sizes for unpublished studies ( $g_{intrusion} = -0.36$ ,  $g_{pretesting} = 0.17$ ) than published ones ( $g_{intrusion} = -0.87$ ,  $g_{pretesting} = 0.36$ ). However, we caution against overinterpreting this result given the relatively small number of unpublished studies ( $k_{intrusion} = 7$ ,  $k_{pretesting} = 4$ ).

### **Participant sample.**

Initial testing potentiated new learning regardless of whether data collection occurred in a laboratory ( $g = 0.46$ ) or online ( $g = 0.35$ ), and the two data collection methods produced similar effect sizes,  $Q = 0.63$ ,  $p = .43$ . This was true for the standard procedure and intrusion subsample as well (see Tables 6 and 7). The pretesting subsample, however, showed that laboratory studies produced a smaller benefit,  $g = 0.25$ , than online studies,  $g = 0.64$ . We note that there were only 10 online studies in this sample, and that they all originate from the same laboratory (Kornell, 2014; Vaughn, Hausman, & Kornell, 2016). As a result, we caution that the difference in the effect sizes between laboratory and online studies might be attributable to other variables such as the specific materials and procedures used. Further research is needed to determine whether participant samples differ in their susceptibility to the pretesting influence.

### **Material type.**

Testing enhanced new learning of both word lists,  $g = 0.68$ , and prose/videos,  $g = 0.77$ , but not paired associates,  $g = 0.04$ . A majority ( $k = 43$  out of 68) of the studies involving paired associates, however, also used the interleaving procedure. The moderating effects of material type disappeared when we considered the data absent these studies, as was the case in the standard procedure and intrusion subsamples, both  $ps = .71$ . In the pretesting subsample, material type had only two levels, and the moderator contributed substantial heterogeneity,  $Q = 16.27$ .

Pretesting produced a greater benefit on new learning for trivia questions,  $g = 0.68$ , compared with paired associates,  $g = 0.15$ .

### **Criterial test format.**

Testing potentiated new learning when the final test was free recall, list cued recall, and MMFR,  $g_s > 0.71$ ; however, this benefit was substantially smaller when the final test involved recognition,  $g = 0.44$ , and item cued recall,  $g = 0.16$ . The smaller effect in recognition is perhaps unsurprising, given the insensitivity of recognition to the influence of prior retrieval (Chan & McDermott, 2007; Darley & Murdock, 1971) and the small sample size for this level of the moderator ( $k = 11$ ). More surprising, however, is the small effect for item cued recall. Once again, a majority ( $k = 45$  out of 78) of these studies also used the interleaving procedure. When we examined the data from the standard procedure subsample, item cued recall demonstrated a significant test-potentiated new learning effect,  $g = 0.65$ , that was similar to other criterial test formats,  $Q = 6.01$ ,  $p = .20$ , and similar patterns were observed for the intrusion and pretesting subsamples (see Tables 7 and 8).

### **Corrective feedback.**

When initial testing was not accompanied by feedback, it substantially facilitated new learning,  $g = 0.73$ . When new learning is equivalent to feedback, as in the single-list pretesting procedure, testing also facilitated new learning,  $g = 0.34$ . Surprisingly, initial testing harmed subsequent learning when feedback was given,  $g = -0.49$ . We suspect, once again, that a design confound between feedback and interleaving might have contributed to this test-impaired learning effect, as 20 of the 23 samples that have administered feedback during initial testing also used the interleaving procedure. Unfortunately, given the high level of confounding between interleaving and feedback, we were unable to examine the effect of feedback on test-potentiated

new learning free from the influence of interleaving – future research is needed to address this issue.

### **Retention interval.**

The beneficial effects of testing on new learning diminished with longer retention intervals,  $B = -0.08$ ,  $p = .03$ . This finding is somewhat surprising given the oft-repeated results in the testing effect literature, whereby the benefits of retrieval practice on the tested, original-learning material (rather than new-learning material) are particularly evident following a substantial (e.g., 24 hrs or more) retention interval. Unlike the voluminous demonstrations in the testing effect literature, only 10 samples here included a retention interval that was at least 24 hr long. The negative association between retention interval and new learning remained even when these 10 samples were removed from the analysis (such that the retention interval ranged from 0 s to 25 min),  $B = -0.13$ ,  $CI [-0.24, -0.03]$ ,  $Z = -2.50$ ,  $p = .01$ . This finding suggests that even a short retention interval can reduce the magnitude of TPNL. Such a conclusion, however, could be premature, given that most existing studies have used a relatively short retention interval, thereby restricting the range of the moderator variable.

### **Evaluating the Theoretical Accounts**

Using the predictions outlined in Tables 8-11, one can assess the theories in the traditional way – i.e., by qualitatively examining whether each moderator result supports a theory's prediction. To facilitate this qualitative assessment, we combined all of the conclusions from the theoretically-motivated moderators section and displayed them graphically in Figure 7. From this perspective, resource theories and integration theories receive considerable support based on the analysis of moderator variables. Indeed, these theories accurately predicted the pattern for a majority (i.e., 4 out of 5) of their relevant moderators. The metacognitive account

did fairly well, too, as it predicted the pattern of three of five variables. In contrast, the context account did rather poorly – it accurately predicted only one out of four variables. Therefore, the qualitative assessment favors resource theories and integration theories.

Another way to assess the theories' predictive power is to conduct a quantitative assessemnt via meta-regression. Note that we conducted the meta-regression analysis for only the complete sample of studies, because the subsamples were too homogenous in their methodology for theoretical testing. The premise of this analysis is that each theory predicts that certain study characteristics should promote test-potentiated new learning relative to other study characteristics; therefore, studies that contain more such characteristics should be associated with a larger TPNL effect than studies that contain fewer such characteristics. To this end, we coded each study for the number of theoretically-derived TPNL characteristics that it contained, and used this score as a predictor for the effect size (Michie et al., 2009; Prestwich et al., 2016). For example, resource theory predicts that blocked presentation, episodic initial testing, having participants study related original- and new-learning items, no delay between original and new learning, and high memory load are methodological characteristics that should promote TPNL (see Table 8 for the rationale of these predictions). If Study A implemented three of these five characteristics, then it would receive a resource theory-derived score of 3. If Study B implemented none of these characteristics, then it would receive a score of 0. Therefore, according to resource theories, Study A should produce a larger TPNL effect than Study B. Because we were testing four theories, and each theory made different predictions about whether or not a given study characteristic would promote TPNL, we computed four theoretically-derived scores for each study: one based on resource theories, one based on metacognitive theories, one based on context theories, and one based on integration theories.

A key assumption of this meta-regression analysis is that we treated all methodological characteristics (i.e., all moderators) with equal importance. Indeed, because the four major classes of explanations are more akin to theoretical frameworks than to formal theories, they do not make specific predictions about which variables might be more important and which might be less so. Consequently, all moderators received equal weights. One might be tempted to suggest that we could use the results from the prior moderator analysis to assign weights, but we opted not to do this for two reasons. First, assigning weights this way is post hoc. Second, and more importantly, because the theories themselves do not make any explicit assumptions about the relative importance of the moderators, assigning weights based on the moderator analysis may actually violate the spirit of the conceptual frameworks. To ensure that all moderators were weighed equally, we assigned a score of 1 to the level(s) of the moderator that is predicted to promote TPNL (e.g., blocked presentation according to resource theory) and a score of 0 – representing baseline – for the remaining level(s) of the moderator (e.g., if that study used an interleaved presentation). Given the current state of the theoretical frameworks, this binary coding scheme satisfactorily characterizes all of the predictions. Future research may help to further solidify these frameworks and allow them to make more fine-grained predictions that go beyond a binary coding scheme (e.g., 0, .5, 1 or -1, 0, 1).

We applied this 0 vs. 1 binary coding scheme to all categorical moderators. To ensure that continuous moderators were weighed in a similar range (i.e., 0 – 1), we performed a logarithmic transformation on the memory load variable (which ordinarily range from 1 to 12) such that it ranged from 0 to 1.08. The retrieval practice variable was already expressed in proportion terms, thereby conveniently aligning with the 0 – 1 scale (with its actual range being .20 - .97). When this variable needed to be reversed coded, as was the case for the

prediction based on metacognitive theories, we subtracted the raw proportion recalled from 1. Lastly, because the moderator of delay between original and new learning spanned a very large range, we converted it into a dichotomous categorical variable, with studies that administered no delay vs. studies that inserted a delay as the two categories.

Coding of the studies are done based on the rationale displayed in Tables 8-11 and the theory-driven scores that each study received are shown in the rightmost columns of the Appendix. To provide an independent evaluation for each class of theories, we conducted four meta-regressions, with the theory-derived score as the predictor and study effect size as the dependent variable.<sup>7</sup> Overall, the results of these meta-regressions (see Table 10) converge with the qualitative analysis shown in Figure 7. Specifically, resource theories and integration theories, the two accounts deemed most successful in the qualitative assessment, accounted for 29% and 30% of the between-study variance, respectively. In contrast, metacognitive theories and context theories accounted for only 6% and 3% of the among-study heterogeneity, respectively.

Next, we examined whether the four theories might be combined to account for additional variance using a simultaneous meta-regression. The model accounted for a combined 44% of the study variance. The results of this analysis are displayed at the bottom of Table 12. Intriguingly, despite accounting for 6% of the variance when considered on its own,

---

<sup>7</sup> One may question why we computed the theory-derived scores by summing the number of methodological characteristics instead of simply inserting the coded moderators into a multiple-regression model. This is because doing the meta-regression in the latter way would not allowed us to take the directionality of the predictions into account even if the moderators were coded for direction (i.e., 0 for baseline and 1 for facilitation). For example, if we have a moderator with two levels, coding level 1 as the baseline condition and level 2 as the facilitation condition would give the same regression results as coding the levels in the opposite direction – both coding schemes will result in this moderator accounting for equal amount of variance in the dependent variable.

metacognitive theories failed to account for a significant amount of unique variance when combined with other theories. Moreover, given that resource theories and integration theories each accounted for approximately 30% of the between-study variance, it is perhaps surprising that the full model accounted for only 44% of the variance. This suppression effect indicates that substantial multicollinearity might have existed in the data. Indeed, an examination of the intercorrelations amongst the four sets of theory-derived scores showed that the resource theory scores and the integration theory scores were highly correlated,  $r = -.70$ , with none of the remaining intercorrelations exceeding .31 (context X integration). When one examines the predictions made by resource theories and integration theories (see Tables 8 and 11), it is perhaps not surprising that their theory-derived scores were correlated. Specifically, these theories made similar predictions for their three shared moderators (i.e., interleaving, relation between original- and new-learning, and initial test format), even though they arrived at those predictions based on different underlying reasonings.

To provide a more definitive conclusion regarding the predictive ability of the four theories – and especially between resource theories and integration theories – we pitted them against each other in a dominance analysis (see Table 13, Azen & Budescu, 2003; Budescu, 1993; Tighe & Schatschneider, 2013). In a dominance analysis, all possible combinations of regression models were conducted and the predictive ability of the variables (and in the present case, resource theories and integration theories) were compared by examining their  $R^2$  when the variables were added to the same regression model. For example, one may compare the predictive power of resource theories vs. integration theories when they were added to a regression model that previously contained only metacognitive theories (row M in Table 13), context theories (row C in Table 13), a combination of metacognitive and context theories (row

“M, C” in Table 13), etc. The results of this dominance analysis clearly favored integration theories over all others, including resource theories. In particular, the predictive power of integration theories exceeded that of resource theories in every individual and averaged pairwise comparisons. In the parlance of dominance analysis, the variable for integration theories thus exerts complete dominance over the variable for resource theories, and is thus considered the better predictor of the two.

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

### **Discussion**

Test-potentiated new learning is a widely reported finding. To provide a comprehensive assessment of the data, we conducted separate moderator analyses for the complete sample of studies, including correct recall data and the subset of studies that reported intrusion data, studies that used only the multi-list, “standard” procedure, and studies that used the single-list, pretesting procedure. Despite its popularity as an empirical phenomenon, theoretical development regarding test-potentiated new learning have mainly been characterized by the introduction of disparate and focused explanations that are relatively narrow in scope. In the present paper, we synthesized these explanations into four classes of accounts and evaluated them with a meta-analysis. Notably, both the qualitative examination and the meta-regression analysis revealed that integration theories and resource theories received considerable support from the data, whereas metacognitive theories and context theories received far more limited support. Below,



we discuss the implications of our results for these theories and then address broader issues of the literature. Finally, we present a study space analysis that highlights research areas that deserve further inquiry.

*Resource theories* suggest that prior retrieval potentiates new learning because it reduces the proactive interference associated with original learning and its detrimental effect on new learning (Malis, 1970), or that it restores the attentional capacity depleted by original learning when one has to learn new material (e.g., Pastötter et al., 2011; Szpunar, 2017). Therefore, manipulations that affect the ability for retrieval to prevent proactive interference from original learning should have an impact on TPNL. In the present data set, moderators that were relevant to resource theories include interleaving, relation between original and new learning, initial test format, memory load, and delay. Overall, resource theories have garnered substantial support from the data, as it correctly predicted the results for four of the five moderators (see the first row of Figure 7). However, despite these generally positive outcomes, resource theories had trouble accounting for situations where testing enhances new learning even when proactive interference is expected to play little to no role. For example, although interleaving reduced TPNL in the multi-list procedure, it did not do the same in the single-list, pretesting procedure. It is not clear how resource theories could address this dissociation, given that the participant-generated target (i.e., the guess from the pretest) is hypothesized to interfere with new learning when the target is presented immediately after, so pretesting should not enhance new learning. More broadly, resource theories appear to have difficulty explaining the benefits of testing in single-list designs (e.g., Kornell, 2014; Experiment 4 of Wissman et al., 2011). Recent findings using the change recollection task (Wahlheim, 2015) also contradicted the idea that prior retrieval can prevent memory of the tested items from intruding during subsequent learning.

Aside from these issues, the idea that testing can increase attention to the study material is generally supported.

*Metacognitive theories* attribute test-potentiated new learning to the optimization of encoding strategy brought about by the metacognitive knowledge that one gains through initial testing. The moderators of interleaving, research design, initial test format, test format match, and retrieval practice performance provide important tests for this account. In general, the metacognitive account has received limited support from the present data – it correctly predicted the patterns for three (i.e., interleaving, research design, and initial test format) of the five relevant moderating variables. But it accounted for virtually no unique variance beyond either resource theories or integration theories. Moreover, conceptually, the biggest weakness of this account is in its inability to explain why a test that is not expected to enhance learners' metacognitive knowledge can still potentiate new learning. For example, nonepisodic initial tests such as semantic generation or N-back should produce no advantage over control tasks, yet these tests have sometimes been effective in facilitating new learning in the multi-list learning paradigm (Divis & Benjamin, 2014; Pastötter et al., 2011; but see also Weinstein et al., 2015). Another prediction emerging from metacognitive theories is that initial testing facilitates later learning by informing participants' expectations regarding the final test (e.g., the test format that they will encounter). Based on this logic, the benefits of retrieval on new-learning should be particularly evident when the initial and final tests share the same format. With the exception of the intrusion data, the present results were largely inconsistent with this prediction (see also Wissman et al., 2011).

Context theories state that performing retrieval practice between study episodes causes an internal context change, which serves to segregate the learning episodes before and after the

retrieval. Learners are then better equipped to retrieve the new-learning information by constraining their search set during the criterial test. Relevant moderators for this account include interleaving, comparison task, relation between original and new learning, and delay between original and new learning. With the exception of interleaving, predictions based on context theories have received little support from the data. In general, context theories have trouble accounting for a TPNL effect that occurs when context segregation is expected to play a limited role (e.g., in the single-list, pretesting procedure) or when the control task should also induce context change (thereby negating the context-changing advantage created by retrieval). For example, a smaller TPNL effect should be observed when the comparison task involved filler activities (e.g., playing a video game, drawing pictures) relative to no-test or restudy (which is presumed not to cause context change). The present results do not support this prediction, with retrieval practice generating similar or greater benefits when compared to filler/math tasks (which should produce a context change for the control participants) than when it was compared to no-test/restudying (which should not produce a context change), and this pattern was observed across all samples. At a theoretical level, the context change idea also suffers from some level of circularity. Specifically, why does retrieval practice, even nonepisodic retrieval practice, enhance context isolation when other distractors tasks (e.g., mental arithmetic) supposedly fail to do so?<sup>8</sup> To date, there is no proven method to examine a priori whether a given task would, or would not,

---

<sup>8</sup> It is not clear why performing mental arithmetic would not cause context change relative to encoding a list of words, given the high degree of differences between the two tasks. In fact, mental arithmetic is frequently used as the baseline, non-context-changing task for studies that investigate context effects in memory (Abel & Bauml, 2016; Sahakyan et al., 2004; Sahakyan & Hendricks, 2012). In a review of the list-based directed forgetting literature (Sahakyan et al., 2013), Sahakyan and colleagues acknowledged that understanding why a given task would, or would not, instill context change is an important topic for future research in the context change literature.

instill internal context change. As a result, when a task fails to produce effects that are hypothesized to be context-dependent, it is difficult to ascertain whether the task fails because it does not initiate a context change or if the hypothesis of context change is incorrect.

*Integration theories* suggest that completing an interpolated test in the multi-list design should increase the likelihood for learners to covertly retrieve the tested, original-learning items during new learning. This covert retrieval is hypothesized to facilitate integration across the different items sets, and the formation of these unified representations should increase their subsequent accessibility. Alternatively, attempting semantic retrieval can activate the memory network that is associated with the retrieval cue in the single-list design, which can in turn facilitate integration between the retrieval cue and the target (i.e., the new-learning item). Integration theories provide testable predictions for the moderators of interleaving, comparison task, item relation, initial test format, and retrieval practice performance. Notably, results from both the qualitative examination and the meta-regression are largely favorable to this account. Indeed, the integration account is the only one that correctly predicted opposites effects of interleaving on TPNL in the multi-list and single-list paradigms (see Table 11 for the rationale), with interleaving reducing TPNL in the former and enhancing it in the latter (see Tables 4 and 7 for the results). Despite the largely positive outcomes, integration theories had trouble accounting for the lack of association between retrieval practice performance and the magnitude of TPNL. Provided that retrieval practice performance serves as an index of original learning, higher retrieval practice performance should be associated with more frequent spontaneous retrievals (and thus integration) during new learning. However, in the present data set, retrieval practice performance was not predictive of new learning (see Figure 5).

### **The State of the Science**

Overall, our meta-analytic results offered the most support for integration theories, with resource theories being a close second. As evidenced by the empty cells in Figure 7, which indicate variables for which an account does not offer a clear prediction, the four accounts are not meant to be direct competitors and are not mutually exclusive – that is, they cover somewhat different aspects of the test-potentiated new learning effect. Given the various forms and manifestations of the test-potentiated new learning phenomenon, it is possible that a combination of factors drive its occurrence and that different mechanisms are responsible for paradigm-specific effects.

It is reasonable to suggest that because these theories tackle somewhat different aspects of the test-potentiated new learning effect, one might combine them to provide a more comprehensive account of TPNL. For example, the basic tenets of resource theories (i.e., testing reduces proactive interference and inattention) and metacognitive theories (i.e., testing of the original-learning material causes participants to alter their encoding strategy for new learning) are quite compatible, so it is entirely possible that testing can facilitate new learning because it increases the attentional resources for new learning while also allowing participants to better use those resources. Alternatively, one might attempt to combine the strengths of context theories and integration theories, as these theories, when combined, produced the greatest increase in variance accounted for (see Table 13). However, attempts at combining any of the four theories must first address the contradictions between them. For example, context theories attribute the benefits of testing on new learning to enhanced segregation between the original and new learning episodes, whereas integration theories claim that testing potentiates new learning because it allows one to combine materials learned across the two episodes. Some researchers

have considered the possibility that initial testing could either segregate or integrate the encoding episodes depending on the learner's goal (e.g., whether one wants to free up cognitive resources or to learn information that builds upon existing knowledge) or the amount of information that one has to learn (Wahlheim, 2015). Alternatively, initial testing might enhance the likelihood that a previously studied episode is spontaneously retrieved during subsequent learning, and this causes learners to encode both pieces of information together while tagging the components of the integrated trace with separate temporal markers (i.e., one may specify which part of the integrated trace came from original learning and which came from new learning, Chan & LaPaglia, 2013). This type of encoding would thus lead to integration while also facilitating source discrimination.

Although the present results favor the resource account and (especially) the integration account, several important theoretical questions remain unanswered. For example, TPNL can be revealed in enhanced correct recall and reduced intrusions in the multi-list design, and this mirror effect is often observed in the literature. However, results from correct recall and intrusions have sometime diverged (e.g., test format match and retrieval practice performance had different effects on correct recall and intrusions, see also Pierce et al., 2017; Weinstein et al., 2014). Further examinations of these dissociations may be particularly illuminating from a theoretical perspective. Another important consideration for future work is to examine whether or not testing facilitates new learning across the various methodological paradigms based on a similar mechanism. Although several theorists have argued that the same processes may underlie all TPNL-like effects (Finn, 2017; Kornell & Vaughn, 2016; Pastötter & Bauml, 2014), as have we, it is also clear that the same moderator variable does not always produce the same results across the subsamples of studies. Investigations into these dissociations may shed light on critical

theoretical mechanisms for TPNL. Lastly, theoretical development regarding TPNL may benefit from research that focuses on questions that allow one to tease apart different theories. For example, as we have alluded to previously, resource theories and integration theories produced similar predictions for several variables, but because of their different underlying logic – resource theories attribute TPNL to retrieval preventing intrusions from original learning into new learning, whereas integration theories attribute TPNL to retrieval increasing these types of “intrusions,” which allow learners to integrate materials across lists – they will invariably make conflicting predictions for certain variables. One possibility is the variable of corrective feedback. When corrective feedback is provided during retrieval practice of the original-learning material, it should increase the likelihood of intrusions during new learning. Therefore, providing feedback should reduce TPNL according to resource theories, whereas it should increase TPNL according to integration theories. In the present data set, we considered feedback as an empirical variable because it was fully confounded with interleaving, but future empirical work can pit resource theories and integration theories against each other by manipulating corrective feedback.

### **Gaps in the Literature**

To help spur future research and to identify gaps in the literature, we present a study space analysis (Malpass et al., 2008) for the correct recall data in Table 14. A study space is essentially a contingency table that displays the frequency with which a particular variable has been investigated. For example, a study space that contains the variables of interleaving (Yes vs. No) and research design (within-subjects vs. between-subjects) would depict the number of effect sizes that exist in each of the 2 X 2 cell. Examination of such a study space can help to guide future studies by identifying under-represented areas of research.

The current study space analysis features the theoretically-relevant variables explored in our meta-analysis. To help visually identify areas of under-representation, we have highlighted variable combinations in which no studies have been conducted (e.g., no studies have compared interleaved testing against interleaved math), and have bolded cells in which the number of studies is less than would be expected by chance. For example, given that there were 159 effect sizes, any cells with fewer than 39 (i.e.,  $159 / 4$ ) effect sizes in a 2 X 2 contingency table would be highlighted, and any cells with fewer than 17 (i.e.,  $159 / 9$ ) effect sizes in a 3 X 3 contingency table would be highlighted. An important exception to this rule is that some variable combinations are not suitable for a specific paradigm. For example, in the single-list, pretesting design, because the initial test requires semantic retrieval (i.e., guessing) and the final test requires episodic retrieval, the variable of test format match (between the initial and criterial tests) would always be “No.” In these cases, we omitted this variable combination when calculating chance frequency. To provide a concrete example, the chance frequency for the study space involving initial test format (which included free recall, item cued recall, list cued recall, nonepisodic recall, and pretest) and test format match was calculated based on a 3 X 2 instead of 5 X 2 design. Because it would not make sense to match the test format when the initial test involved nonepisodic retrieval, the nonepisodic recall and pretest levels were dropped from the chance frequency calculation.

We opted to compare actual frequencies with expected frequencies based on an even distribution instead of chi-squared test of independence because the latter takes into account frequency differences across levels in a main effect. For example, in the 2 X 2 (retention interval: < 1 day vs. > 1 day) frequency table featuring research design (within- vs. between-subjects) and retention interval (< 1 day, > 1 day) as independent factors, only between-subjects studies with >



1 day retention interval ( $k = 3$ ) would be classified as under-investigated. In contrast, within-subjects studies with a retention interval  $> 1$  day ( $k = 5$ ) would not be classified as under-investigated. We feel that this is problematic given that it is clear that more studies with  $> 1$  day retention interval are needed in this literature regardless of research design.

An examination of the study space revealed several areas that have received little attention. For example, very few studies ( $k = 6$ ) implemented a *> 1 day delay* between original learning and new learning. Moreover, aside from the pretesting paradigm, very few studies used *nonepisodic retrieval* as the initial test format ( $k = 8$ ), and only 11 used *recognition* in the criterial test. In fact, *recognition has not yet been used as the initial test format* for any studies in the present sample. Such variables are clearly deserving of further investigation.

This analysis also provides a way to identify confounding in the study space, which can be achieved by locating cells with unusually high frequencies within a particular level of a moderating variable. For example, given that interleaving/blocking had a such powerful impact on test-potentiated new learning, one can examine whether a high concentration of studies that used the interleaving procedure might have been confounded with another moderator. This can be seen in the first column of the Table 11, which shows the frequency distributions of studies that used the interleaving procedure. While the moderator of interleaving was not confounded with research design, because interleaving studies were relatively evenly distributed across the two levels of this moderator ( $k = 30$  for within-subjects and  $k = 21$  for between-subjects), *interleaving was confounded with comparison task*, as demonstrated by the very high concentration of interleaving studies within the restudy level of the comparison task moderator. Future research should examine whether *interleaving* would impair new learning to the same degree when it is compared to a *filler task* instead of restudy.

An examination for areas of confounding also highlights areas that need further investigation. For example, *feedback* is an important variable from an educational perspective. If instructors insert brief quizzes during a lecture, it is likely that students will receive corrective feedback on those questions. An examination of Table 14 reveals that amongst the 23 studies that have implemented feedback in the multi-list design, only three were conducted in the blocked (i.e., standard) paradigm.

While it is beyond the scope of the present paper to investigate every possible combination of the moderator variables for confounding, the presence of confounding or interactions based upon uneven study distributions (e.g., interleaving studies tended to use item-cued recall for retrieval practice) is a natural and common occurrence among any sizable literature. Our objective of presenting this study space is to provide a means for the interested readers to identify areas of future investigations that interest them.

### **Concluding Remarks**

Performing memory retrieval before an encoding task can potentiate new learning. This technique, whether implemented as an interpolated test or a pretest, has been proposed as a method to optimize learning in the classroom, particularly in situations that demand learners to sustain their attention for an extended period of time (e.g., classroom lecture, employee training sessions, tutoring sessions, online webinars). In this paper, we provided an integrative review of the theoretical explanations that have been offered to explain findings related to the phenomenon of test-potentiated new learning, and we organized these explanations into the categories of resource accounts, metacognitive accounts, context accounts, and integration accounts. In addition, we produced a quantitative summary of the extant empirical data, and used these results to show that integration theories have garnered the strongest support.

Future research should continue to examine the influence of under-studied variables as revealed by the study space analysis (e.g., what are the effects of delay or feedback on new learning?), disentangle factors that have been confounded in the literature (e.g., most of the studies that used the interleaving procedure also used paired-associates as the study materials), further understand the boundary conditions for test-potentiated new learning (e.g., when does the testing impair new learning?), and attempt to elucidate and refine the existing theoretical accounts by tackling important questions (e.g., why and when does nonepisodic recall enhance new learning?). As we have shown in the present analysis, testing can facilitate future learning across a diverse array of conditions, and advances in our understanding of this phenomenon offer significant potential for optimizing learning in (and out of) the classroom.

### References

- Abel, M., & Bauml, K. H. (2016). Retrieval practice can eliminate list method directed forgetting. *Memory & Cognition*, 44, 15–23.
- Allen, G., & Arbak, C. J. (1976). The priority effect in the A-B, A-C paradigm and subjects' expectations. *Journal of Verbal Learning and Verbal Behavior*, 15, 381–385.
- Arkes, H. R., & Lyons, D. J. (1979). A mediational explanation of the priority effect. *Journal of Verbal Learning and Verbal Behavior*, 18, 721–731.
- Arnold, K. M., & McDermott, K. B. (2013a). Free recall enhances subsequent learning. *Psychonomic Bulletin & Review*, 20, 507–513.
- Arnold, K. M., & McDermott, K. B. (2013b). Test-potentiated learning: Distinguishing between direct and indirect effects of tests. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 39, 940–945.
- Aslan, A., & Bauml, K. H. (2016). Testing enhances subsequent learning in older but not in younger elementary school children. *Developmental Science*, 19, 992–998.
- Azen, R., & Budescu, D. V. (2003). The dominance analysis approach for comparing predictors in multiple regression. *Psychological Methods*, 8, 129–148.
- Balota, D. A., & Neely, J. H. (1980). Test-expectancy and word-frequency effects in recall and recognition. *Journal of Experimental Psychology: Human Learning and Memory*, 6, 576–587.
- Barnes, J. M., & Underwood, B. J. (1959). “Fate” of first-list associations in transfer theory. *Journal of Experimental Psychology*, 58, 97–105.
- Bauml, K. H., & Kliegl, O. (2013). The critical role of retrieval processes in release from proactive interference. *Journal of Memory and Language*, 68, 39–53.

- Bertsch, S., Pesta, B. J., Wiscott, R., & McDaniel, M. A. (2007). The generation effect: a meta-analytic review. *Memory & Cognition*, 35, 201–210.
- Bettencourt, K., Delaney, P. F., & Chang, Y. (2015). The relationship between test-impaired learning and the testing effect. *Poster Presented at the Annual Meeting of the North Carolina Cognition Group*. Elon, North Carolina.
- Brewer, G. A., Marsh, R. L., Meeks, J. T., Clark-Foos, A., & Hicks, J. L. (2010). The effects of free recall testing on subsequent source memory. *Memory*, 18, 385–393.
- Briggs, G. E. (1954). Acquisition, extinction, and recovery functions in retroactive inhibition. *Journal of Experimental Psychology*, 47, 285–293.
- Budescu, D. V. (1993). Dominance Analysis: A New Approach to the Problem of Relative Importance of Predictors in Multiple Regression. *Psychological Bulletin*, 114, 542–551.
- Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk: A New Source of Inexpensive, Yet High-Quality, Data? *Perspectives on Psychological Science*, 6, 3–5.
- 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, 1547–1552.
- Chan, J. C. K., & LaPaglia, J. A. (2011). The dark side of testing memory: repeated retrieval can enhance eyewitness suggestibility. *Journal of Experimental Psychology: Applied*, 17, 418–432.
- Chan, J. C. K., & LaPaglia, J. A. (2013). Impairing existing declarative memory in humans by disrupting reconsolidation. *Proceedings of the National Academy of Sciences of the United States of America*, 110, 9309–9313.

- Chan, J. C. K., & McDermott, K. B. (2007). The testing effect in recognition memory: A dual process account. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 33, 431–437.
- Chan, J. C. K., Erdman, M. R., & Davis, S. D. (2015). Retrieval induces forgetting, but only when nontested items compete for retrieval: Implication for interference, inhibition, and context reinstatement. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 41, 1298–1315.
- Chan, J. C. K., Manley, K. D., & Lang, K. (2017). Retrieval-enhanced suggestibility: A retrospective and a new investigation. *Journal of Applied Research in Memory and Cognition*, 6, 213–229.
- Chan, J. C. K., Manley, K. D., Davis, S. D., & Szpunar, K. K. (2018). Testing potentiates new learning across a retention interval and a lag: A strategy change perspective.
- 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, 553–571.
- Chan, J. C. K., Thomas, A. K., & Bulevich, J. B. (2009). Recalling a witnessed event increases eyewitness suggestibility: The reversed testing effect. *Psychological Science*, 20, 66–73.
- Chan, J. C. K., Wilford, M. M., & Hughes, K. L. (2012). Retrieval can increase or decrease suggestibility depending on how memory is tested: The importance of source complexity. *Journal of Memory and Language*, 67, 78–85.
- Cho, K. W., Neely, J. H., Crocco, S., & Vitrano, D. (2017). Testing Enhances Both Encoding and Retrieval for Both Tested and Untested Items. *The Quarterly Journal of Experimental Psychology*, 70, 1211–1235.

- Congleton, A., & Rajaram, S. (2011). The origin of the interaction between learning method and delay in the testing effect: The roles of processing and conceptual retrieval organization. *Memory & Cognition*, 40, 528–539.
- Cumming, G. (2012). *Understanding The New Statistics*. New York, NY: Routledge.
- Darley, C. F., & Murdock, B. B. (1971). Effects of prior free recall testing on final recall and recognition. *Journal of Experimental Psychology*, 91, 66–73.
- Davis, S. D., & Chan, J. C. K. (2015). Studying on borrowed time: how does testing impair new learning? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 41, 1741–1754.
- Davis, S. D., Chan, J. C. K., & Wilford, M. M. (2017). The dark side of interpolated testing: Frequent switching between retrieval and encoding impairs new learning. *Journal of Applied Research in Memory and Cognition*, 6, 434–441.
- Divis, K., & Benjamin, A. S. (2014). Retrieval speeds context fluctuation: Why semantic generation enhances later learning but hinders prior learning. *Memory & Cognition*, 42, 1049–1062.
- Donaldson, W. (1971). Output effects in multitrial free recall. *Journal of Verbal Learning and Verbal Behavior*, 10, 577–585.
- Dunlosky, J., & Ariel, R. (2011). Self-regulated learning and the allocation of study time. In *Psychology of Learning and Motivation* (Vol. 54, pp. 103–140). Elsevier.
- Dunlosky, J., & Connor, L. (1997). Age-related differences in the allocation of study time account for age-related differences in memory performance. *Memory & Cognition*, 25, 691–700.

- Finley, J. R., & Benjamin, A. S. (2012). Adaptive and qualitative changes in encoding strategy with experience: evidence from the test-expectancy paradigm. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 38, 632–652.
- Finn, B. (2017). A Framework of Episodic Updating: An Account of Memory Updating After Retrieval. In *Psychology of Learning and Motivation* (Vol. 67, pp. 173–211). Elsevier.
- Finn, B., & Roediger, H. L. (2013). Interfering effects of retrieval in learning new information. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 39, 1665–1681.
- Godden, D., & Baddeley, A. (1975). Context-dependent memory in two natural environments: on land and underwater. *British Journal of Psychology*, 66, 325–331.
- Goodman, J. K., Cryder, C. E., & Cheema, A. (2013). Data collection in a flat world: the strengths and weaknesses of Mechanical Turk samples. *Journal of Behavioral Decision Making*, 26, 213–224.
- Gordon, L. T. (2016). *Retrieval and attention: Factors that potentiate learning and enhance eyewitness suggestibility*. Unpublished dissertation.
- Gordon, L. T., & Thomas, A. K. (2014). Testing potentiates new learning in the misinformation paradigm. *Memory & Cognition*, 42, 186–197.
- Gordon, L. T., Thomas, A. K., & Bulevich, J. B. (2015). Looking for answers in all the wrong places: How testing facilitates learning of misinformation. *Journal of Memory and Language*, 83, 140–151.
- Grimaldi, P. J., & Karpicke, J. D. (2012). When and why do retrieval attempts enhance subsequent encoding? *Memory & Cognition*, 40, 505–513.



- Gunter, B. (1980). Release from proactive interference with television news items: Evidence for encoding dimensions within televised news. *Journal of Experimental Psychology: Human Learning and Memory*, 6, 216–223.
- Hays, M. J., Kornell, N., & Bjork, R. A. (2013). When and why a failed test potentiates the effectiveness of subsequent study. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 39, 290–296.
- Hintzman, D. L. (2004). Judgment of frequency versus recognition confidence: Repetition and recursive reminding. *Memory & Cognition*, 32, 336–350.
- Hintzman, D. L. (2009). How does repetition affect memory? Evidence from judgments of recency. *Memory & Cognition*, 38, 102–115.
- Howard, M. W., & Kahana, M. J. (2002). A distributed representation of temporal context. *Journal of Mathematical Psychology*, 46, 269–299.
- Huff, M. J., Davis, S. D., & Meade, M. L. (2013). The effects of initial testing on false recall and false recognition in the social contagion of memory paradigm. *Memory & Cognition*, 41, 820–831.
- Huff, M., Bodner, G. E., & Fawcett, J. M. (2015). Effects of distinctive encoding on correct and false memory: A meta-analytic review of costs and benefits and their origins in the DRM paradigm. *Psychonomic Bulletin & Review*, 22, 349–365.
- Hupbach, A., Gomez, R., & Nadel, L. (2009). Episodic memory reconsolidation: Updating or source confusion? *Memory*, 17, 502–510.
- Hupbach, A., Gomez, R., & Nadel, L. (2011). Episodic memory updating: The role of context familiarity. *Psychonomic Bulletin & Review*, 18, 787–797.

Hupbach, A., Gomez, R., Hardt, O., & Nadel, L. (2007). Reconsolidation of episodic memories:

A subtle reminder triggers integration of new information. *Learning & Memory*, 14, 47–53.

Izawa, C. (1967). Function of test trials in paired-associate learning. *Journal of Experimental*

*Psychology*, 75, 194–209.

Izawa, C. (1970). Optimal potentiating effects and forgetting-prevention effects of tests in

paired-associate learning. *Journal of Experimental Psychology*, 83, 340–344.

Izawa, C. (1971). The test trial potentiating model. *Journal of Mathematical Psychology*, 8, 200–

224.

Jacoby, L. L., Shimizu, Y., Daniels, K. A., & Rhodes, M. G. (2005). Modes of cognitive control

in recognition and source memory: Depth of retrieval. *Psychonomic Bulletin & Review*, 12,

852–857.

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, 1282–1297.

Jang, Y., & Huber, D. E. (2008). Context retrieval and context change in free recall: Recalling

from long-term memory drives list isolation. *Journal of Experimental Psychology: Learning,*

*Memory, and Cognition*, 34, 112–127.

Jensen, O., Gelfand, J., Kounios, J., & Lisman, J. E. (2002). Oscillations in the alpha band (9-12

Hz) increase with memory load during retention in a short-term memory task. *Cerebral*

*Cortex*, 12, 877–882.

Jing, H. G., Szpunar, K. K., & Schacter, D. L. (2016a). Interpolated testing influences focused

attention and improves integration of information during a video-recorded lecture. *Journal of*

*Experimental Psychology: Applied*, in press, 1–47.

- Jing, H. G., Szpunar, K. K., & Schacter, D. L. (2016b). Interpolated testing influences focused attention and improves integration of information during a video-recorded lecture. *Journal of Experimental Psychology: Applied*, 22, 305–318.
- Johnstone, A. H., & Percival, F. (1976). Attention Breaks in Lectures. *Education in Chemistry*, 13, 49–50.
- Jonker, T. R., Seli, P., & MacLeod, C. M. (2013). Putting retrieval-induced forgetting in context: An inhibition-free, context-based account. *Psychological Review*, 120, 852–872.
- Jonker, T. R., Seli, P., & MacLeod, C. M. (2015). Retrieval-induced forgetting and context. *Current Directions in Psychological Science*.
- Kang, S. H. K. (2010). Enhancing visuospatial learning: the benefit of retrieval practice. *Memory & Cognition*, 38, 1009–1017.
- Karpicke, J. D. (2009). Metacognitive control and strategy selection: Deciding to practice retrieval during learning. *Journal of Experimental Psychology: General*, 138, 469–486.
- Klein, K., & Boals, A. (2001). Expressive writing can increase working memory capacity. *Journal of Experimental Psychology: General*, 130, 520–533.
- Kliegl, O., Pastötter, B., & Bauml, K. H. (2015). The contribution of encoding and retrieval processes to proactive interference. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 41, 1778–1789.
- Kornell, N. (2014). Attempting to answer a meaningful question enhances subsequent learning even when feedback is delayed. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 40, 106–114.
- Kornell, N., & Vaughn, K. E. (2016). How Retrieval Attempts Affect Learning (Vol. 65, pp. 183–215). Elsevier.

- Kornell, N., Hays, M. J., & Bjork, R. A. (2009). Unsuccessful retrieval attempts enhance subsequent learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 35, 989–998.
- Lane, S. M., Mather, M., Villa, D., & Morita, S. (2001). How events are reviewed matters: Effects of varied focus on eyewitness suggestibility. *Memory & Cognition*, 29, 940–947.
- LaPaglia, J. A., & Chan, J. C. K. (2013). Testing increases suggestibility for narrative-based misinformation but reduces suggestibility for question-based misinformation. *Behavioral Sciences and the Law*, 31, 593–606.
- LaPaglia, J. A., Wilford, M. M., Rivard, J. R., Chan, J. C. K., & Fisher, R. P. (2014). Misleading suggestions can alter later memory reports even following a Cognitive Interview. *Applied Cognitive Psychology*, 28, 1–9.
- Lau, J. (2006). The case of the misleading funnel plot. *British Medical Journal*, 333, 597–600.
- Lee, H. S., & Ahn, D. (2017). Testing Prepares Students to Learn Better: The Forward Effect of Testing in Category Learning. *Journal of Educational Psychology*. doi:10.1037/edu0000211
- Lee, J. L. C., Nader, K., & Schiller, D. (2017). An Update on Memory Reconsolidation Updating. *Trends in Cognitive Sciences*, 0, 531–545.
- Lehman, M., Smith, M. A., & Karpicke, J. D. (2014). Toward an episodic context account of retrieval-based learning: dissociating retrieval practice and elaboration. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 40, 1787–1794.
- Little, J. L., & McDaniel, M. A. (2015). Metamemory monitoring and control following retrieval practice for text. *Memory & Cognition*, 43, 85–98.

- MacLeod, C. M., Gopie, N., Hourihan, K. L., Neary, K. R., & Ozubko, J. D. (2010). The production effect: delineation of a phenomenon. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 36, 671–685.
- Malis, G. (1970). *The priority effect in the A-B, A-C paradigm*. Unpublished dissertation.
- Malpass, R. S., Tredoux, C. G., Compo, N. S., McQuiston Surret, D., MacLin, O. H., Zimmerman, L. A., & Topp, L. D. (2008). Study space analysis for policy development. *Applied Cognitive Psychology*, 22, 789–801.
- May, R. B., & Thompson, J. M. (1989). Test expectancy and question answering in prose processing. *Applied Cognitive Psychology*, 3, 261–269.
- Mayer, R. E., DeLeeuw, K. E., & Ayres, P. (2007). Creating retroactive and proactive interference in multimedia learning. *Applied Cognitive Psychology*, 21, 795–809.
- Mayer, R. E., Mayer, R. E., Stull, A., DeLeeuw, K., Almeroth, K., Bimber, B., et al. (2009). Clickers in college classrooms: Fostering learning with questioning methods in large lecture classes. *Contemporary Educational Psychology*, 34, 51–57.
- Michie, S., Abraham, C., Whittington, C., McAteer, J., & Gupta, S. (2009). Effective techniques in healthy eating and physical activity interventions: A meta-regression. *Health Psychology*, 28, 690–701.
- Morris, C. D., Bransford, J. D., & Franks, J. J. (1977). Levels of processing versus transfer appropriate processing. *Journal of Verbal Learning and Verbal Behavior*, 16, 519–533.
- Nelson, S. M., Arnold, K. M., Gilmore, A. W., & McDermott, K. B. (2013). Neural signatures of test-potentiated learning in parietal cortex. *Journal of Neuroscience*, 33, 11754–11762.
- Nunes, L. D., & Weinstein, Y. (2012). Testing improves true recall and protects against the build-up of proactive interference without increasing false recall. *Memory*, 20, 138–154.

- Palva, S., & Palva, J. M. (2007). New vistas for  $\alpha$ -frequency band oscillations. *Trends in Neurosciences*, 30, 150–158.
- Paolacci, G. (2010). Running experiments on Amazon mechanical turk. *Judgment and Decision Making*, 5, 411–419.
- Park, D., Ramirez, G., & Beilock, S. L. (2014). The role of expressive writing in math anxiety. *Journal of Experimental Psychology: Applied*, 20, 103–111.
- Pashler, H., Kang, S. H. K., & Mozer, M. C. (2013). Reviewing erroneous information facilitates memory updating. *Cognition*, 128, 424–430.
- Pastötter, B., & Bauml, K. H. (2014). Retrieval practice enhances new learning: the forward effect of testing. *Frontiers in Psychology*, 5, 1–5.
- Pastötter, B., & Bauml, K. H. (2016). Untitled. *Unpublished Data*.
- Pastötter, B., Schicker, S., Niedernhuber, J., & Bauml, K. H. (2011). Retrieval during learning facilitates subsequent memory encoding. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 37, 287–297.
- Pastötter, B., Weber, J., & Bauml, K. H. (2013). Using testing to improve learning after severe traumatic brain injury. *Neuropsychology*, 27, 280–285.
- Pennebaker, J. W., & Chung, C. K. (2011). Expressive Writing: Connections to Physical and Mental Health. In H. S. Friedman, *The Oxford Handbook of Health Psychology* (pp. 417–437). New York, NY: Oxford University Press.
- Picho, K., Rodriguez, A., & Finnie, L. (2013). Exploring the moderating role of context on the mathematics performance of females under stereotype threat: a meta-analysis. *The Journal of Social Psychology*, 153, 299–333.

- Pierce, B. H., & Hawthorne, M. J. (2016). Does the testing effect depend on presentation modality? *Journal of Applied Research in Memory and Cognition*, 5, 52–58.
- Pierce, B. H., Gallo, D. A., & McCain, J. L. (2017). Reduced interference from memory testing: a postretrieval monitoring account. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, in press. doi:10.1037/xlm0000377
- Postman, L. (1961). The present status of interference theory. In C. N. Cofer, *Verbal Learning and Verbal Behavior* (pp. 152–179). New York, NY.
- Potts, R., & Shanks, D. R. (2012). Can testing immunize memories against interference? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 38, 1780–1785.
- Potts, R., & Shanks, D. R. (2014). The benefit of generating errors during learning. *Journal of Experimental Psychology: General*, 143, 644–667.
- Prestwich, A., Kellar, I., Conner, M., Lawton, R., Gardner, P., & Turgut, L. (2016). Does changing social influence engender changes in alcohol intake? A meta-analysis. *Journal of Consulting and Clinical Psychology*, 84, 845–860.
- Pyc, M. A., & Rawson, K. A. (2010). Why testing improves memory: mediator effectiveness hypothesis. *Science*, 330, 335–335.
- Ramirez, G., & Beilock, S. L. (2011). Writing about testing worries boosts exam performance in the classroom. *Science*, 331, 211–213.
- Richland, L. E., Kornell, N., & Kao, L. S. (2009). The pretesting effect: Do unsuccessful retrieval attempts enhance learning? *Journal of Experimental Psychology: Applied*, 15, 243–257.
- Robbins, D., & Irvin, J. R. (1976). The priority effect: Test effects on negative transfer and control lists. *Bulletin of the Psychonomic Society*, 8, 167–168.

- Roets, A., Van Hiel, A., & Cornelis, I. (2006). The dimensional structure of the need for cognitive closure scale: relationships with “seizing” and ‘freezing’ processes. *Social Cognition, 24*, 22–45.
- Rowland, C. A. (2014). The effect of testing versus restudy on retention: a meta-analytic review of the testing effect. *Psychological Bulletin, 140*, 1432–1463.
- Sahakyan, L., & Hendricks, H. E. (2012). Context change and retrieval difficulty in the list-before-last paradigm. *Memory & Cognition, 40*, 844–860.
- Sahakyan, L., Delaney, P. F., Foster, N. L., & Abushanab, B. (2013). List-method directed forgetting in cognitive and Clinical Research. In *Psychology of Learning and Motivation* (Vol. 59, pp. 131–189). Elsevier.
- Sahakyan, L., Delaney, P. F., & Kelley, C. M. (2004). Self-evaluation as a moderating factor of strategy change in directed forgetting benefits. *Psychonomic Bulletin & Review, 11*, 131–136.
- Schacter, D. L. (2001). Forgotten ideas, neglected pioneers: Richard Semon and the story of memory. Philadelphia: Psychology Press.
- Schacter, D. L., Eich, J. E., & Tulving, E. (1978). Richard Semon's theory of memory. *Journal of Verbal Learning and Verbal Behavior, 17*, 721–743.
- Scully, I. D., Napper, L. E., & Hupbach, A. (2016). Does reactivation trigger episodic memory change? A meta-analysis. *Neurobiology of Learning and Memory*.  
doi:10.1016/j.nlm.2016.12.012
- Shimizu, Y., & Jacoby, L. L. (2005). Similarity-guided depth of retrieval: Constraining at the front end. *Canadian Journal of Experimental Psychology, 59*, 17–21.



- Simons, D. J., Alogna, V. K., Attaya, M. K., Aucoin, P., Bahník, Š., Birch, S., et al. (2014). Registered replication report: Schooler and Engstler-Schooler (1990). *Perspectives on Psychological Science*, 9, 556–578.
- Smallwood, J., & Schooler, J. W. (2015). The science of mind wandering: empirically navigating the stream of consciousness. *Annual Review of Psychology*, 66, 487–518.
- Soderstrom, N. C., & Bjork, R. A. (2014). Testing facilitates the regulation of subsequent study time. *Journal of Memory and Language*, 73, 99–115.
- Son, L. K., & Kornell, N. (2008). Research on the allocation of study time: Key studies from 1890 to the present (and beyond). In J. Dunlosky & R. A. Bjork, *Handbook of metamemory and memory* (pp. 333–351). New York: Psychology Press.
- St Jacques, P. L., Olm, C., & Schacter, D. L. (2013). Neural mechanisms of reactivation-induced updating that enhance and distort memory. *Proceedings of the National Academy of Sciences of the United States of America*, 110, 19671–19678.
- Stuart, J., & Rutherford, R. J. D. (1978). Medical student concentration during lectures. *Lancet*, 312, 514–516.
- Szpunar, K. (2017). Directing the Wandering Mind. *Current Directions in Psychological Science*, 26, 40–44.
- Szpunar, K. K., Jing, H. G., & Schacter, D. L. (2014). Overcoming overconfidence in learning from video-recorded lectures: Implications of interpolated testing for online education. *Journal of Applied Research in Memory and Cognition*, 3, 161–164.
- Szpunar, K. K., Khan, N. Y., & Schacter, D. L. (2013). Interpolated memory tests reduce mind wandering and improve learning of online lectures. *Proceedings of the National Academy of Sciences of the United States of America*, 110, 6313–6317.

- Szpunar, K. K., McDermott, K. B., & Roediger, H. L. (2007). Expectation of a final cumulative test enhances long-term retention. *Memory & Cognition*, 35, 1007–1013.
- Szpunar, K. K., McDermott, K. B., & Roediger, H. L. (2008). Testing during study insulates against the buildup of proactive interference. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 34, 1392–1399.
- Terrin, N., Schmid, C. H., & Lau, J. (2005). In an empirical evaluation of the funnel plot, researchers could not visually identify publication bias. *Journal of Clinical Epidemiology*, 58, 894–901.
- Thomas, A. K., Bulevich, J. B., & Chan, J. C. K. (2010). Testing promotes eyewitness accuracy with a warning: Implications for retrieval enhanced suggestibility. *Journal of Memory and Language*, 63, 149–157.
- Thomas, R. C., & McDaniel, M. A. (2013). Testing and Feedback Effects on Front-End Control Over Later Retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 39, 437–450.
- Tighe, E. L., & Schatschneider, C. (2013). A dominance analysis approach to determining predictor importance in third, seventh, and tenth grade reading comprehension skills. *Reading and Writing*, 27, 101–127.
- Topp, N. W., & Pawloski, B. (2002). Online data collection. *Journal of Science Education and Technology*, 11, 173–178.
- Tulving, E. (1983). *Elements of Episodic Memory*. New York: Oxford University Press.
- Tulving, E., & Pearlstone, Z. (1966). Availability versus accessibility of information in memory for words. *Journal of Verbal Learning and Verbal Behavior*, 5, 381–391.

- Tulving, E., & Thompson, D. M. (1973). Encoding specificity and retrieval processes in episodic memory. *Psychological Review*, 80, 352–373.
- Tulving, E., & Thomson, D. M. (1973). Encoding specificity and retrieval processes in episodic memory. *Psychological Review*, 80, 352–373.
- Tulving, E., & Watkins, M. J. (1974). On negative transfer: Effects of testing one list on the recall of another. *Journal of Verbal Learning and Verbal Behavior*, 13, 181–193.
- Tulving, E., Patterson, R. D., & Malis, G. (1967). Unpublished manuscript.
- Underwood, B. J. (1948). “Spontaneous recovery” of verbal associations. *Journal of Experimental Psychology*, 38, 429–439.
- van Kesteren, M. T. R., Brown, T. I., & Wagner, A. D. (2016). Interactions between Memory and New Learning: Insights from fMRI Multivoxel Pattern Analysis. *Frontiers in Systems Neuroscience*, 10, 1–5.
- Vaughn, K. E., & Rawson, K. A. (2012). When is guessing incorrectly better than studying for enhancing memory? *Psychonomic Bulletin & Review*, 19, 899–905.
- Vaughn, K. E., Hausman, H., & Kornell, N. (2016). Retrieval attempts enhance learning regardless of time spent trying to retrieve. *Memory*, *in press*, 1–20.
- Verkoeijen, P. P. J. L., Tabbers, H. K., & Verhage, M. L. (2011). Comparing the effects of testing and restudying on recollection in recognition memory. *Experimental Psychology (Formerly Zeitschrift Für Experimentelle Psychologie)*, 58, 490–498.
- Wahlheim, C. N. (2015). Testing can counteract proactive interference by integrating competing information. *Memory & Cognition*, 43, 27–38.
- Wahlheim, C. N., & Jacoby, L. L. (2010). Experience with proactive interference diminishes its effects: mechanisms of change. *Memory & Cognition*, 39, 185–195.

- Webster, D. M., & Kruglanski, A. W. (1994). Individual differences in need for cognitive closure. *Journal of Personality and Social Psychology*, 67, 1049–1062.
- Weinstein, Y., Gilmore, A. W., Szpunar, K. K., & McDermott, K. B. (2014). The role of test expectancy in the build-up of proactive interference in long-term memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 40, 1039–1048.
- Weinstein, Y., McDermott, K. B., & Szpunar, K. K. (2011). Testing protects against proactive interference in face–name learning. *Psychonomic Bulletin & Review*, 18, 518–523.
- Weinstein, Y., McDermott, K. B., Szpunar, K. K., Bauml, K. H., & Pastötter, B. (2015). Not all retrieval during learning facilitates subsequent memory encoding(?). Presented at the the annual meeting of the Psychonomic Society, Chicago, IL.
- Weinstein, Y., Nunes, L. D., & Karpicke, J. D. (2016). On the placement of practice questions during study. *Journal of Experimental Psychology: Applied*, 22, 72–84.
- Whiffen, J. W., & Karpicke, J. D. (2017). The Role of Episodic Context in Retrieval Practice Effects. *Journal of Experimental Psychology: Learning, Memory, and Cognition*.  
doi:10.1037/xlm0000379
- Wickens, D. D. (1970). Encoding categories of words: An empirical approach to meaning. *Psychological Review*, 77, 1–15.
- Wilford, M. M., Chan, J. C. K., & Tuhn, S. J. (2014). Retrieval enhances eyewitness suggestibility to misinformation in free and cued recall. *Journal of Experimental Psychology: Applied*, 20, 81–93.
- Wissman, K. T., & Rawson, K. A. (2015a). Grain size of recall practice for lengthy text material. *Unpublished Data*.

- Wissman, K. T., & Rawson, K. A. (2015b). Grain size of recall practice for lengthy text material: fragile and mysterious effects on memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*, 439–455.
- 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*, 1140–1147.
- Young, J. L. (1971). Reinforcement-test intervals in paired-associate learning. *Journal of Mathematical Psychology*, *8*, 58–81.
- Yue, C. L., Soderstrom, N. C., & Bjork, E. L. (2015). Partial testing can potentiate learning of tested and untested material from multimedia lessons. *Journal of Educational Psychology*, 1–16.
- Zaromb, F. M., & Roediger, H. L. (2010). The testing effect in free recall is associated with enhanced organizational processes. *Memory & Cognition*, *38*, 995–1008.

**Table 1***Search terms used in PsycINFO and the number of results returned*

Search Terms	Number of Results
test* AND “proactive interference”	457
test* AND “new learning”	420
test* AND (“subsequent study” OR “later study” OR “subsequent learning” OR “later learning”)	373
“immediate test*”	303
(retrieval practice) AND (new learning)	237
“inter* test*”	215
misinformation AND source AND test	74
(retrieval practice) AND (proactive interference)	17
“test-potentiate*”	13
“priority effect”	16
"interpolated testing" OR "intervening testing" OR "interim testing"	17
misinformation AND “initial test*”	12
“test* effect” AND misinformation	8
“repeated testing” AND “new learning”	2

## Table 2

*Stem-and-leaf plot of the effect sizes based on correct recall data, with the median effect size highlighted by an outside border*

Stem	Leaf (Correct Recall Data)									
1.9	5									
1.8										
1.7										
1.6	0	3	4	9						
1.5	3									
1.4	2	4	5	8						
1.3	0	2	4	8						
1.2	0	5	7	8	8	9	9			
1.1	5	9								
1.0	2	2	2	2	2	5	6	6	7	8
0.9	0	1	1	3	3	4	5	6	7	7
0.8	0	0	0	1	2	3	4	5	6	
0.7	0	3	5	5	5	6	6	8		
0.6	1	1	3	3	5	6	6	7	9	9
0.5	0	0	0	2	3	7	7	8	9	9
0.4	0	4	4	5	5	7	7	7	7	
0.3	1	3	4	4	6	6	7	7	9	9
0.2	1	3	6	7	9					
0.1	0	1	3	5	6	7				
0.0	0	0	5	5	6	6	7	7		
-0.0	7									
-0.1	0	4	5	6	8					
-0.2	0	3	7	9						
-0.3	2	4	5							
-0.4	0	1	3	7	9					
-0.5	1	1	1	3	5	5	6			
-0.6	4	8								
-0.7	4									
-0.8	1	1	8							
-0.9	4									
-1.0										
-1.1										
-1.2										
-1.3										
-1.4										
-1.5										
-1.6	2	4	8							

*Note.* Effect sizes are displayed in a descending order, such that larger benefits of retrieval practice on new learning are shown towards the top of the table. Negative effect sizes indicate that retrieval practiced impaired new learning. Studies that used the standard procedure are shown in the darker-shaded cells, and studies that used the pretesting procedure are shown in the lighter-shaded cells.



**Table 3**

*Stem and leaf plot based on the intrusion data, with the median effect size highlighted by an outside border*

Stem	Leaf (Intrusion Data)			
-3.0	0			
-2.9				
-2.8				
-2.7				
-2.6				
-2.5				
-2.4				
-2.3				
-2.2				
-2.1				
-2.0				
-1.9				
-1.8				
-1.7	0			
-1.6	1			
-1.5	1			
-1.4				
-1.3	8	9		
-1.2	3	4	5	9
-1.1	3	4		
-1.0	2			
-0.9	3	6	8	
-0.8	1	6	7	7
-0.7	0	1	4	9
-0.6	9			
-0.5	0	3	4	6 9
-0.4	2	7		
-0.3				
-0.2	9			
-0.1	2	6	8	9
-0.0	7			
0.0	0	1	5	

*Note.* Because reduction of intrusions based on retrieval practice is shown as a negative effect size, the effect sizes here are display in an ascending order, such that larger benefits of retrieval practice on new learning are again shown towards the top of the table.

**Table 4**

*Influence of moderators on test-potentiated new learning based on correct recall data in the complete sample*

Moderator	Point Estimate	.95 CI	$Q_B$	Z	p	k
Overall Effect	0.44	[0.35, 0.53]	1132.84	9.42	< .01	159
Interleaved Testing with New Learning			56.91		< .01	
No	0.66	[0.56, 0.77]		12.65	< .01	108
Yes	-0.02	[-0.16, 0.13]		-0.24	.81	51
Research Design			10.09		< .01	
Between-Subjects	0.55	[0.44, 0.66]		9.55	< .01	105
Within-Subjects	0.25	[0.11, 0.40]		3.41	< .01	54
Comparison Task			49.35		< .01	
Filler	0.79	[0.53, 1.05]		5.90	< .01	18
Math	0.74	[0.53, 0.95]		6.85	< .01	28
No Test	0.64	[0.49, 0.79]		8.20	< .01	47
Restudy*	0.09	[-0.04, 0.22]		1.39	.17	66
Relation between Original and New Learning			15.09		< .01	
Unrelated	0.25	[0.13, 0.38]		3.91	< .01	78
Related	0.60	[0.48, 0.73]		9.69	< .01	81
Initial test Format			36.05		< .01	
Free Recall	0.72	[0.44, 0.99]		5.06	< .01	14
Item Cued Recall*	0.17	[0.02, 0.32]		2.25	.02	54
List Cued Recall	0.82	[0.65, 1.00]		9.16	< .01	38
Nonepisodic Recall	0.32	[-0.09, 0.73]		1.55	.12	8
Pretest	0.31	[0.19, 0.50]		4.43	< .01	45
Test Format Match			2.96		.09	
No	0.50	[0.38, 0.61]		8.57	< .01	105

Moderator	<i>Point Estimate</i>	<i>.95 CI</i>	<i>Q<sub>B</sub></i>	<i>Z</i>	<i>p</i>	<i>k</i>
Yes	0.33	[0.18, 0.48]		4.19	< .01	54
Memory Load	0.08	[0.03, 0.13]		2.93	< .01	159
Retrieval Practice Performance*	1.17	[0.43, 1.92]		3.10	< .01	78
Delay Between Original and New Learning	-0.05	[-0.11, 0.02]		-1.27	.20	159
Publication Status			7.47		.01	
No	0.19	[-.00, 0.39]		1.93	.05	33
Yes	0.50	[0.40, 0.60]		9.69	< .01	126
Participant Sample			0.57		.45	
In Lab	0.45	[0.35, 0.55]		8.99	< .01	138
Online	0.35	[0.12, 0.59]		2.90	< .01	21
Material Type			78.36		< .01	
Words	0.66	[0.50, 0.82]		8.17	< .01	38
Paired Associates	0.04	[-0.08, 0.15]		0.64	.52	68
Prose/Videos	0.77	[0.64, 0.90]		11.60	< .01	53
Criterion Test Format			47.50		< .01	
Free Recall	0.88	[0.65, 1.10]		7.55	< .01	21
Item Cued Recall*	0.16	[0.04, 0.28]		2.65	< .01	78
List Cued Recall	0.69	[0.51, 0.87]		7.45	< .01	36
MMFR	0.76	[0.45, 1.06]		4.86	< .01	13
Recognition	0.44	[0.13, 0.76]		2.76	< .01	11
Administration of Corrective Feedback			110.16		< .01	
No	0.72	[0.62, 0.82]		13.57	< .01	91
Yes* - Multi-list Paradigm	-0.50	[-0.70, -0.29]		-4.75	< .01	23
Yes - Pretesting Paradigm	0.34	[0.21, 0.48]		4.92	< .01	45

Moderator	<i>Point Estimate</i>	<i>.95 CI</i>	$Q_B$	$Z$	$p$	$k$
Retention Interval	-0.08	[-0.15, -0.01]		-2.18	.03	159

---

*Note:* Asterisks next to moderator names indicate that the moderator effect might have been driven by a confounding factor or by inclusion of outlying data points.

**Table 5**

*Influence of moderators on test-potentiated new learning based on correct recall data in the standard procedure subsample*

Moderator	Point Estimate	.95 CI	$Q_B$	Z	p	k
Overall Effect	0.75	[0.66, 0.85]	280.93	15.36	< .01	84
Research Design			23.03		< .01	
Between-Subjects	0.84	[0.75, 0.93]		17.73	< .01	71
Within-Subjects	0.34	[0.15, 0.52]		3.60	< .01	13
Comparison Task			2.32		.51	
Filler	0.81	[0.58, 1.03]		6.90	< .01	17
Math	0.84	[0.65, 1.04]		8.33	< .01	21
No Test	0.74	[0.59, 0.88]		11.99	< .01	34
Restudy	0.61	[0.35, 0.86]		4.61	< .01	12
Relation between Original and New Learning			1.33		.25	
Unrelated	0.69	[0.55, 0.84]		9.47	< .01	39
Related	0.81	[0.68, 0.94]		12.09	< .01	45
Initial test Format			0.87		.65	
Free Recall	0.72	[0.49, 0.94]		6.28	< .01	14
Item Cued Recall	0.71	[0.55, 0.87]		8.57	< .01	32
List Cued Recall	0.81	[0.66, 0.95]		11.00	< .01	38
Test Format Match			4.30		.04	
No	0.86	[0.72, 1.00]		12.20	< .01	45
Yes	0.66	[0.53, 0.79]		9.82	< .01	39
Memory Load*	-0.01	[-0.06, 0.04]		-0.36	.72	84
Retrieval Practice Performance	0.15	[-0.57, 0.86]		0.41	.68	56
Delay Between Original and New Learning*	-0.10	[-0.17, -0.03]		-2.89	< .01	84
Publication Status			< 0.01		.93	
No	0.76	[0.55, 0.97]		7.12	< .01	16
Yes	0.75	[0.64, 0.86]		13.50	< .01	68
Participant Sample			0.12		.73	
In Lab	0.76	[0.66, 0.86]		14.79	< .01	79
Online	0.69	[0.33, 1.06]		3.74	< .01	5

Moderator	<i>Point Estimate</i>	<i>.95 CI</i>	$Q_B$	$Z$	$p$	$k$
Material Type			0.69		.71	
Words	0.74	[0.58, 0.90]		9.32	< .01	32
Paired Associates	0.69	[0.46, 0.91]		5.92	< .01	17
Prose/Videos	0.80	[0.65, 0.94]		10.61	< .01	35
Criterion Test Format			6.01		.20	
Free Recall	0.88	[0.70, 1.06]		9.54	< .01	21
Item Cued Recall	0.65	[0.44, 0.87]		6.06	< .01	16
List Cued Recall	0.78	[0.62, 0.94]		9.62	< .01	30
MMFR	0.71	[0.47, 0.96]		5.77	< .01	13
Recognition	0.38	[-0.02, 0.79]		1.86	.06	4
Retention Interval*	-0.07	[-0.14, 0.01]		-1.80	.07	84

*Note:* Asterisks next to moderator names indicate that the moderator effect might have been driven by a confounding factor or by inclusion of outlying data points. Interleaving was removed as a moderator from this subsample because all studies in the “standard procedure” used the blocked design. Feedback was omitted as a moderator because it was administered in only three studies in this subsample.

**Table 6***Influence of moderators on test-potentiated new learning based on intrusion data*

Moderator	Point Estimate	.95 CI	$Q_B$	Z	p	k
Overall Effect	-0.77	[-0.93, -0.62]	162.97	-9.62	< .01	41
Research Design			4.28		.04	
Between-Subjects	-0.84	[-1.00, -0.68]		-10.01	< .01	36
Within-Subjects	-0.41	[-0.78, -0.04]		-2.16	.03	5
Comparison Task			1.04		.79	
Filler	-0.89	[-1.37, -0.41]		-3.61	< .01	5
Math	-0.80	[-1.02, -0.58]		-7.14	< .01	22
No Test	-0.80	[-1.20, -0.40]		-5.35	< .01	6
Restudy	-0.61	[-0.98, -0.25]		-3.31	< .01	8
Relation between Original and New Learning			1.76		.19	
Unrelated	-0.66	[-0.89, -0.43]		-5.59	< .01	19
Related	-0.87	[-1.09, -0.66]		-7.92	< .01	22
Initial test Format			4.86		.09	
Item Cued Recall	-0.63	[-0.97, -0.30]		-3.67	< .01	9
List Cued/Free Recall	-0.88	[-1.07, -0.69]		-9.05	< .01	28
Nonepisodic Recall	-0.31	[-0.83, 0.21]		-1.17	.24	4
Test Format Match			5.63		.02	
No	-0.45	[-0.75, -0.14]		-2.85	< .01	10
Yes	-0.87	[-1.04, -0.70]		-10.00	< .01	31
Memory Load*	< -0.01	[-0.09, 0.00]		0.05	.96	41
Retrieval Practice Performance	0.65	[-0.28, 1.58]		1.37	.17	31
Delay Between Original and New Learning*	0.39*	[0.16, 0.61]		3.39	< .01	41
Publication Status			6.24		.01	
No	-0.36	[-0.72, 0.00]		-1.96	.05	7
Yes	-0.87	[-1.04, -0.69]		-9.91	< .01	34
Participant Sample			0.55		.46	
In Lab	-0.75	[-0.92, -0.58]		-8.78	< .01	37

Moderator	<i>Point Estimate</i>	<i>.95 CI</i>	$Q_B$	$Z$	$p$	$k$
Online	-0.93	[-1.38, -0.48]		-4.07	< .01	4
Material Type			0.14		.71	
Words	-0.75	[-0.95, -0.56]		-7.53	< .01	28
Prose/Videos	-0.82	[-1.10, -0.54]		-5.76	< .01	13
Criterion Test Format			0.39		.53	
Item Cued Recall	-0.67	[-1.02, -0.33]		-3.81	< .01	8
List Cued Recall	-0.80	[-0.98, -0.62]		-8.83	< .01	25
Retention Interval	0.08	[-0.07, 0.22]		1.04	.30	41

*Note:* Asterisks next to moderator names indicate that the moderator effect might have been driven by a confounding factor or by inclusion of outlying data points. Interleaving and feedback were omitted as moderators because there were not enough samples to fill each level of these variables.



**Table 7**

*Influence of moderators on test-potentiated new learning based on correct recall data in the single-list, pretesting procedure subsample*

Moderator	Point Estimate	.95 CI	$Q_B$	Z	p	k
Overall Effect	0.35	[0.20, 0.50]	361.29	4.50	< .01	45
Interleaved Testing with New Learning			0.23		.63	
No	0.29	[0.04, 0.55]		2.24	.03	16
Yes	0.37	[0.18, 0.56]		3.87	< .01	29
Research Design			1.65		.20	
Between-Subjects	0.56	[0.20, 0.91]		3.07	< .01	9
Within-Subjects	0.30	[0.13, 0.46]		3.54	< .01	36
Comparison Task			0.03		.85	
No Test	0.37	[0.08, 0.66]		2.50	.01	13
Restudy	0.34	[0.16, 0.52]		3.68	< .01	32
Relation between Original and New Learning			2.60		.11	
Unrelated	0.16	[-0.12, 0.43]		1.12	.26	13
Related	0.42	[0.25, 0.60]		4.71	< .01	32
Delay Between Original and New Learning	-0.16	[-0.28, -0.04]		-2.65	.01	45
Publication Status			0.56		.45	
No	0.17	[-0.33, 0.66]		0.67	.51	4
Yes	0.36	[0.20, 0.52]		4.46	< .01	41
Participant Sample			5.77		.02	
In Lab	0.25	[0.10, 0.41]		3.17	< .01	35
Online	0.64	[0.36, 0.91]		4.53	< .01	10
Material Type			16.27		< .01	
Paired associates	0.15	[-0.01, 0.30]		1.84	.07	29
Trivia questions	0.68	[0.48, 0.89]		6.44	< .01	16
Criterion Test Format			0.11		.74	
Item Cued Recall	0.34	[0.17, 0.50]		4.04	< .01	39

Moderator	<i>Point Estimate</i>	<i>.95 CI</i>	$Q_B$	$Z$	$p$	$k$
Recognition	0.41	[-0.00, 0.83]		1.94	.05	6
Retention Interval	0.05	[-0.04, 0.15]		1.12	.27	45

*Note:* For studies that used the pretesting design, we defined relation between original and new learning based on whether the pretest question was associated with the target (i.e., new learning). Studies in this subsample used pretest as the initial test, administered feedback as new learning, and the initial pretest did not involve episodic retrieval; therefore, initial test format, feedback administration, test format match, and retrieval practice performance were dropped as moderators

**Table 8**

*Study characteristics that should promote test-potentiated new learning according to resource theories*

	Theory-derived TPNL characteristics	Rationale based on Resource Theories
Interleaving	Blocked presentation	In the interleaving procedure, the new-learning trial occurs immediately after retrieval (or restudy) of the original-learning item; consequently, retrieval practice is unlikely to prevent intrusion of the original-learning item during new learning. Therefore, interleaved presentation should weaken test-potentiated new learning, and the opposite is true for blocked presentation.
Relation between original and new learning	Related	Because related materials produce greater proactive interference than unrelated materials (Gunter, 1980; Wickens, 1970), the interference-reducing power of testing should be more evident when participants study the former.
Initial test format	Item cued recall, list cued recall, free recall	Episodic recall tests are required to provide cognitive closure, reducing the need for learners to hold the original-learning items in mind during new learning.
Memory load	Continuous variable – raw number of item sets studied was logarithmic transformed, with higher memory load predicted to produce larger a TPNL effect	The benefits of retrieval on new learning should be particularly evident under greater (rather than less) memory load, when participants in the control condition should experience more proactive interference and mind wandering (Kliegl, Pastötter, & Bauml, 2015; Mayer, DeLeeuw, & Ayres, 2007).
Delay between original and new learning	No delay	Delaying new learning (during which participants are not studying additional materials) should help participants in the control condition restore their attentional capacity, thereby eliminating the advantage conferred by testing. Consequently, the advantage of testing should be maximized when new learning immediately follows original learning.

**Table 9**

*Study characteristics that should promote test-potentiated new learning according to metacognitive theories*

	Theory-derived TPNL Characteristics	Rationale based on Metacognitive Theories
Interleaving	Blocked presentations should increase TPNL for studies that used episodic initial retrieval in the multi-list design. For single-list design, both interleaved and blocked presentations should produce TPNL.	Interleaving original- and new-learning trials can bias learners towards rehearsing the original-learning items in lieu of new learning (Davis & Chan, 2015). Therefore, interleaved presentation should reduce TPNL, and the opposite is true for blocked presentation. Note, however, this bias towards relearning should not occur when initial testing is nonepisodic, so interleaving should not reduce TPNL in the single-list procedure.
Research design	Between-subjects	The benefits of retrieval should be greater when testing is manipulated in a between-subjects design than in a within-subjects design. In the latter case, learners might be able to apply the more efficient encoding strategy that they gained from prior testing to the control condition.
Initial test format	Item cued recall, list cued recall, free recall, pretest	Episodic retrieval tasks are required to facilitate new learning in the multi-list design because only they can provide feedback to the learners about their prior learning. In the single-list design, taking the pretest can inform participants about the nature of cue-target relationship, leading to a shift in encoding strategy.
Test format match	Match	The encoding strategy that learners develop based upon the initial test should be best applied to a similar criterial test (Morris et al., 1977; Tulving & Thomson, 1973).
Retrieval practice performance	Lower performance = Greater TPNL	When participants perform poorly during the initial test, they may be more motivated to learn subsequent materials. Therefore, poorer retrieval practice performance should be associated with a greater TPNL effect.

**Table 10**

*Study characteristics that should promote test-potentiated new learning according to context theories*

	Theory-derived TPNL characteristics	Rationale based on Context Theories
Interleaving	Blocked presentation	Intermixing the retrieval of original-learning items with the encoding of new-learning items should render it very difficult to establish distinct contexts between these items. Consequently, interleaved presentation should eliminate the test-potentiated new learning effect, and blocked presentation should do the opposite.
Comparison task	Restudy, no-test, math	Restudy, no-test, and math are hypothesized not to induce context change in the control condition, thus maximizing the context change advantage conferred by testing when compared to the control condition. In contrast, filler tasks (e.g., drawing pictures, playing a video game) should induce context change from encoding, thereby reducing the advantage of testing on new-learning in comparison.
Relation between original and new learning	Unrelated	When the new-learning items are related to the original-learning items (e.g., they are both four-legged animals), presentation of the new-learning items may remind participants of the original-learning context. Such reminders should weaken the context-isolating benefit of testing and reduce TPNL, a disadvantage that does not apply to unrelated materials.
Delay between original and new learning	No delay	A longer delay between original and new learning should offer the opportunity for a context change even without retrieval, thereby reducing the TPNL effect. In contrast, the context changing power of retrieval should be more evident when there is no delay between original and new learning.

**Table 11**

*Study characteristics that should promote test-potentiated new learning according to integration theories*

	Theory-derived TPNL Characteristics	Rationale based on Integration Theories
Interleaving	Blocked presentation for multi-list studies, interleaved presentation for single-list, pretesting studies	In the multi-list design, testing is hypothesized to potentiate new learning by promoting spontaneous retrieval of the original-learning item when one encodes the new item, thereby facilitating integration. Interleaving should allow integration to occur in both the testing and restudy conditions (because the new-learning item would be presented immediately following retrieval/restudy of the original-item), thus eliminating the advantage of testing over restudying. In the single-list design, pretesting is hypothesized to enhance new learning because attempting semantic retrieval (e.g., whale - ?) primes the network associated with the cue, which in turn facilitates integration of the new-learning target (e.g., mammal) with the cue. In this situation, interleaving is predicted to strengthen TPNL.
Comparison task	Filler, no-test, math	The impact of interpolated testing on new learning should be weaker when it is compared to restudy than when it is compared to other control tasks such as math or filler activities. Because restudy strengthens accessibility of the original-learning items (similar to retrieval practice), it should increase the likelihood of integration in the restudy control condition.
Relation between original and new learning	Related	When the original and new learning items are related, studying the new items should remind participants of the original-learning items, particularly when participants had performed retrieval practice on the original items, facilitating their integration.
Initial test format	Item cued recall, list cued recall, free recall, pretest	Episodic recall tasks should increase spontaneous retrieval of the original-learning items during new learning, thereby facilitating TPNL. Moreover, pretesting should prime the semantic network for new learning and enhances integration of the new-learning material into the just-activated semantic network.
Retrieval practice performance	Higher performance = Greater TPNL	Higher retrieval practice performance should be associated more frequent spontaneous retrieval of the original-learning items during new-learning, and thus a stronger test-potentiated new learning effect.

**Table 12***Results of the meta-regression analysis*

Theory	<i>B</i>	.95 <i>CI</i>	<i>Z</i>	<i>p</i> -value	<i>r</i> <sup>2</sup>
Individual Regression Models					
Resource theories	0.35	[0.27, 0.42]	8.82	< .01	.29
Metacognitive theories	0.15	[0.05, 0.24]	3.12	< .01	.06
Context theories	0.22	[0.08, 0.37]	3.02	< .01	.03
Integration theories	0.36	[0.29, 0.44]	9.27	< .01	.30
Simultaneous Regression Model					.44
Resource theories	0.12	[0.01, 0.23]	2.17	.03	
Metacognitive theories	-0.06	[-0.14, 0.03]	-1.36	.17	
Context theories	0.31	[0.19, 0.43]	4.95	< .01	
Integration theories	0.32	[0.21, 0.44]	5.73	< .01	

*Note: Data in the bottom half of the table denote unique contributions of each theory in the simultaneous regression model.*

**Table 13***Results of the dominance analysis*

Variable(s)	$R^2$	+R	+M	+C	+I
		.292	.060	.033	.302
R	.292	-	-.005*	.045	.043
M	.060	.227	-	.024	.234
C	.033	.303	.051	-	.388
I	.302	.034	-.007*	.120	-
Model Size = 1, Average		.188	.013	.063	.222
R, M	.287	-	-	.049	.046
R, C	.336	-	-.001*	-	.094
R, I	.335	-	-.003*	.095	-
M, C	.084	.252	-	-	.336
M, I	.294	.038	-	.126	-
C, I	.421	.009	-.001*	-	-
Model Size = 2, Average		.100	-.002	.090	.159
R, M, C	.336	-	-	-	.103
R, M, I	.332	-	-	.106	-
R, C, I	.430	-	.008*	-	-
M, C, I	.420	.019	-	-	-
Model Size = 3, Average		.019	.008	.106	.103
Overall Average	.439	.147	.013	.075	.193

*Note: R refers to the variable of resource theory scores, M refers to the variable of metacognitive theory scores, C refers to the variable of context theory scores, and I refers to the variable of integration theory scores. Columns 1 and 2 show the variables in the regression model and their resulting  $R^2$ . Columns 4-6 show the incremental  $R^2$  added by the addition of the variable listed at the top. Asterisks denote conditions in which inclusion of the variable failed to significantly improve the model's fit.*

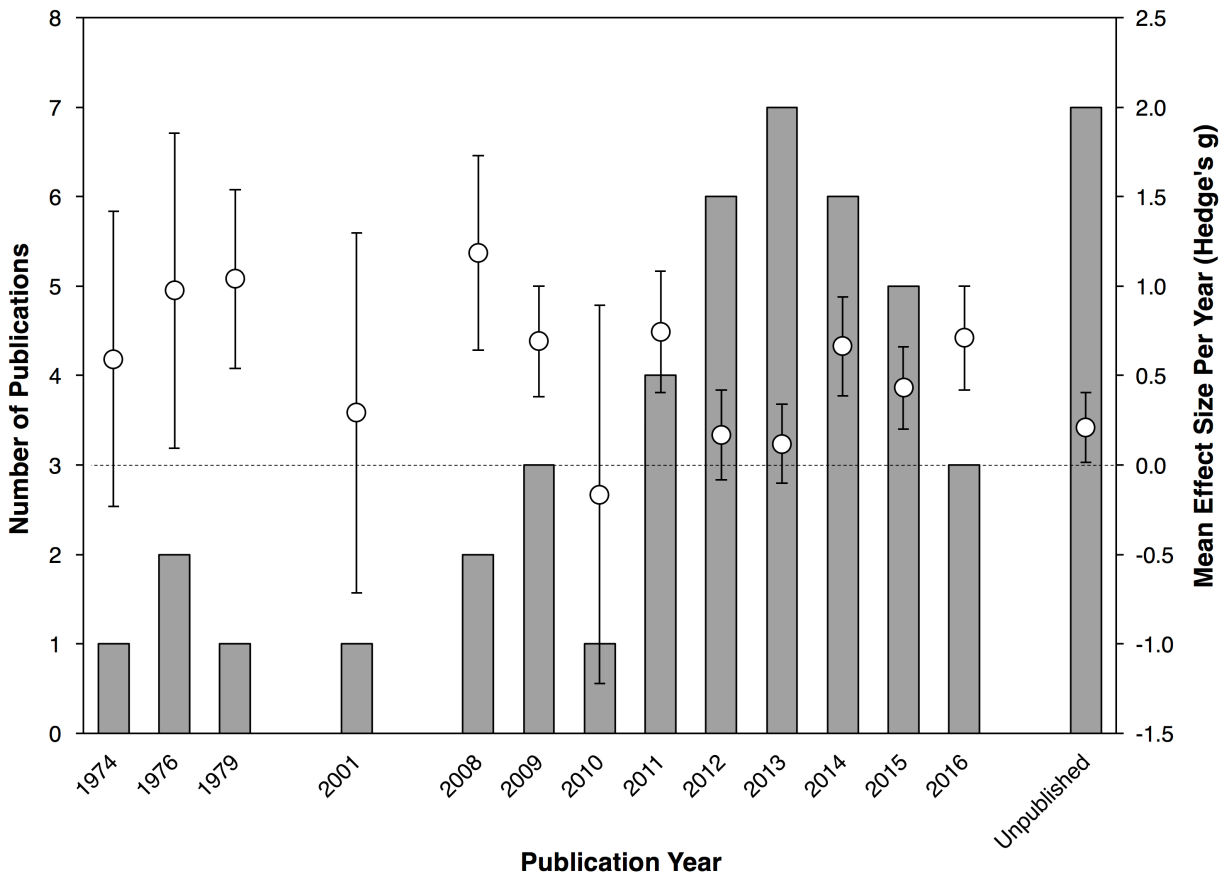


**Table 14**

*Contingency table that displays the study space for moderators of theoretical interest.*

		Interleaving Testing/New Learning		Research Design		Comparison Task				Relation Between Original-New Learning		Initial Test Format					Criterial Test Format					Test Format Match		Feedback		
		Yes (Interleaved)	No (Blocked)	Within-Subjects	Between-Subjects	Restudy	No-Test	Filler	Math	Yes	No	Free Recall	Item Cued Recall	List Cued Recall	Nonepisodic Recall	Pretest	Free Recall	Item Cued Recall	List Cued Recall	MMFR	Recognition	Yes	No	Yes	No	Yes (Pretesting)
Research Design	Within-Subjects	30	24																							
	Between-Subjects	21	84																							
Comparison Task	Restudy	44	22	34	32																					
	No-Test	7	40	15	32																					
	Filler	0	18	4	14																					
	Math	0	28	1	27																					
Relation Between Original-New Learning	Yes	18	63	29	52	25	30	6	20																	
	No	33	45	25	53	41	17	12	8																	
Initial Test Format	Free Recall		14	0	14	0	13	1	0	5	9															
	Item Cued Recall	22	32	10	44	31	7	11	5	14	40															
	List Cued Recall		38	7	31	3	14	5	16	28	10															
	Nonepisodic Recall	0	8	1	7	0	0	1	7	2	6															
	Pretest	29	16	36	9	32	13	0	0	32	13															
Criterial Test Format	Free Recall	0	21	0	21	0	20	0	1	13	8	12	0	9	0	0										
	Item Cued Recall	45	33	38	40	54	17	0	7	40	38	0	36	2	1	39										
	List Cued Recall	0	36	6	30	6	5	7	18	18	18	1	4	25	6											
	MMFR	0	13	4	9	0	5	8	0	6	7	0	12	1												
	Recognition	6	5	6	5	6	0	0	2	4	7	1	2	1	1	6										
Test Format Match	Yes	29	53	32	50	39	22	7	14	36	46	11	25	18			11	41	21	5	4					
	No	22	55	22	55	27	25	11	14	45	32	3	29	20	8	45	3	29	20	8	45					
Feedback	Yes	20	3	5	18	22	0	0	1	2	21	0	22	1	0		0	22	0	0		1	16	7		
	No	2	89	13	78	12	34	18	27	47	44	14	32	37	8		21	17	36	13		4	45	46		
	Yes (Pretesting)	29	16	36	9	32	13	0	0	32	13					45	0	39	0			6	21	24		
Delay Original-New Learning	< 1 Day	51	102	48	105	66	41	18	28	75	78	14	51	36	8	44	21	76	36	9	11	78	75	23	86	44
	> 1 Day		6	6	0	0	6	0	0	6	0	0	3	2	0	1	0	2	0	4	0	4	2	0	5	1

*Note:* Bold text show frequencies that were lower than expected by chance, and shaded cells show areas that have yet to be investigated. Empty cells refer to variable combinations that cannot/should not occur.

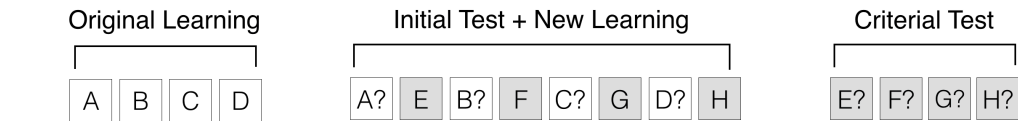


**Figure 1.** Number of publications and magnitude as a function of publication year. The left ordinate and the gray-colored bars display the number of publications per year. The number of unpublished papers are estimated by the sets of studies that appear to belong together. The right ordinate and the data points with .95 *CI* error bars display the average effect size per publication year. The dotted horizontal line indicates an effect size of 0. Because data collection for the meta-analysis ended on April 15, 2016, we grouped studies that are in press at the time of writing into 2016.



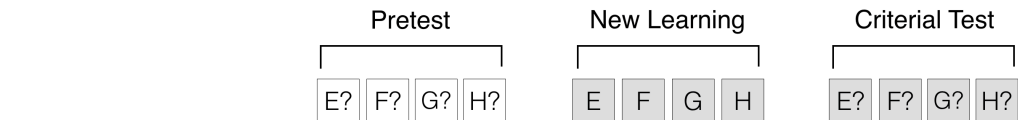
### a) Multi Lists, Blocked Design

---



### b) Multi Lists, Interleaved Design

---



### c) Single List, Blocked Design

---

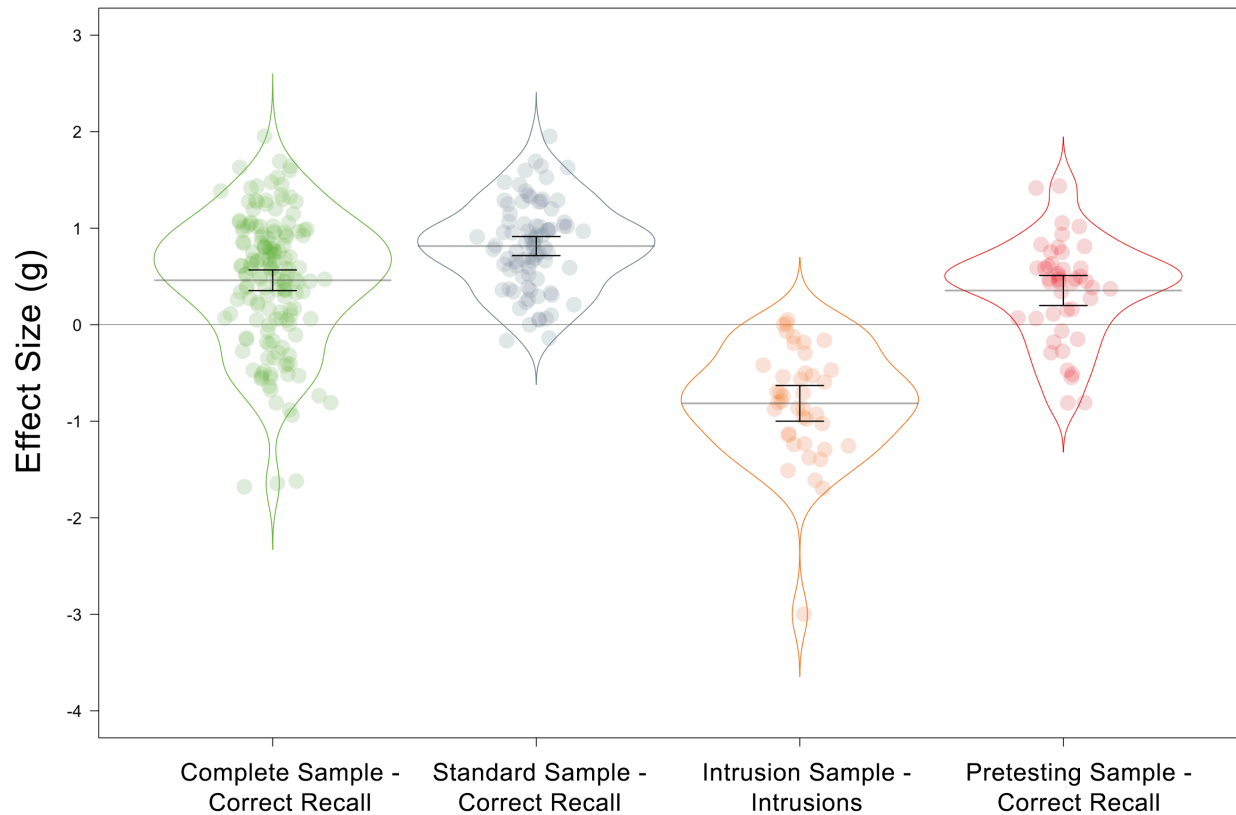


### d) Single List, Interleaved Design

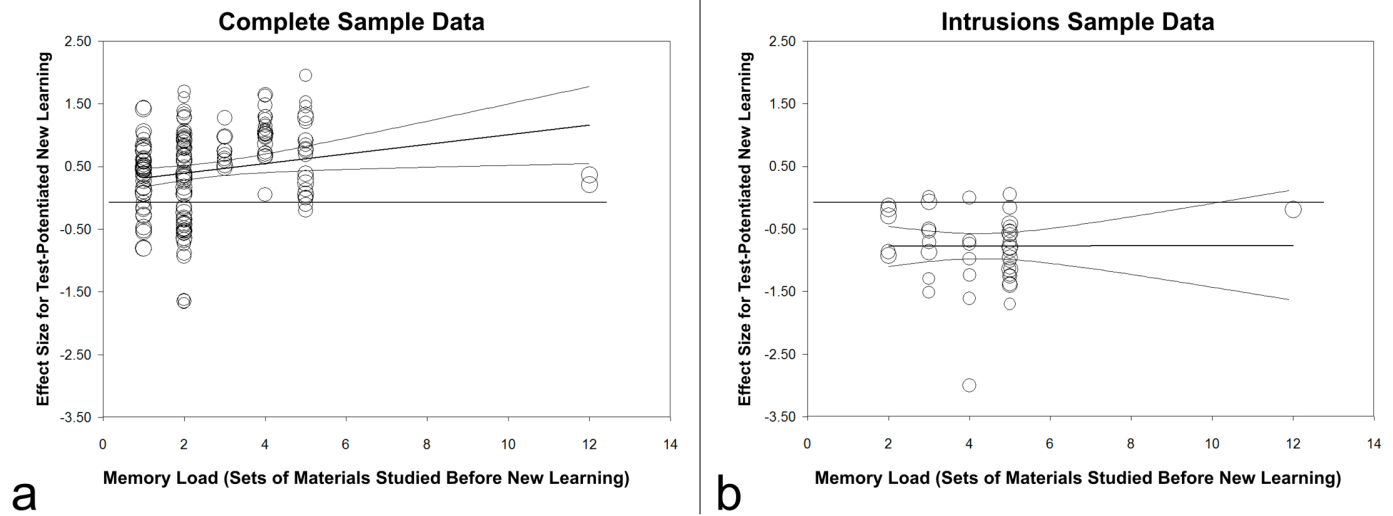
---

**Figure 2.** Paradigms used to investigate the influence of prior retrieval on subsequent learning of new information. Figure 1a shows the most commonly used design, in which participants study some original learning materials, complete an initial test on those materials, and then learn new materials. Figure 1b shows a variant of this design (the interleaved design) in which initial testing and new learning trials are intermixed inside a single block of trials. Figures 1c and 1d shows a single list design in which the original learning phase is omitted. Instead, participants are tested before they learn the new materials. Note that in the single list paradigm, the materials were chosen such that participants would not be able to guess the identity of the to-be-learned item during the initial pretest. Consequently, the item presented for encoding following the

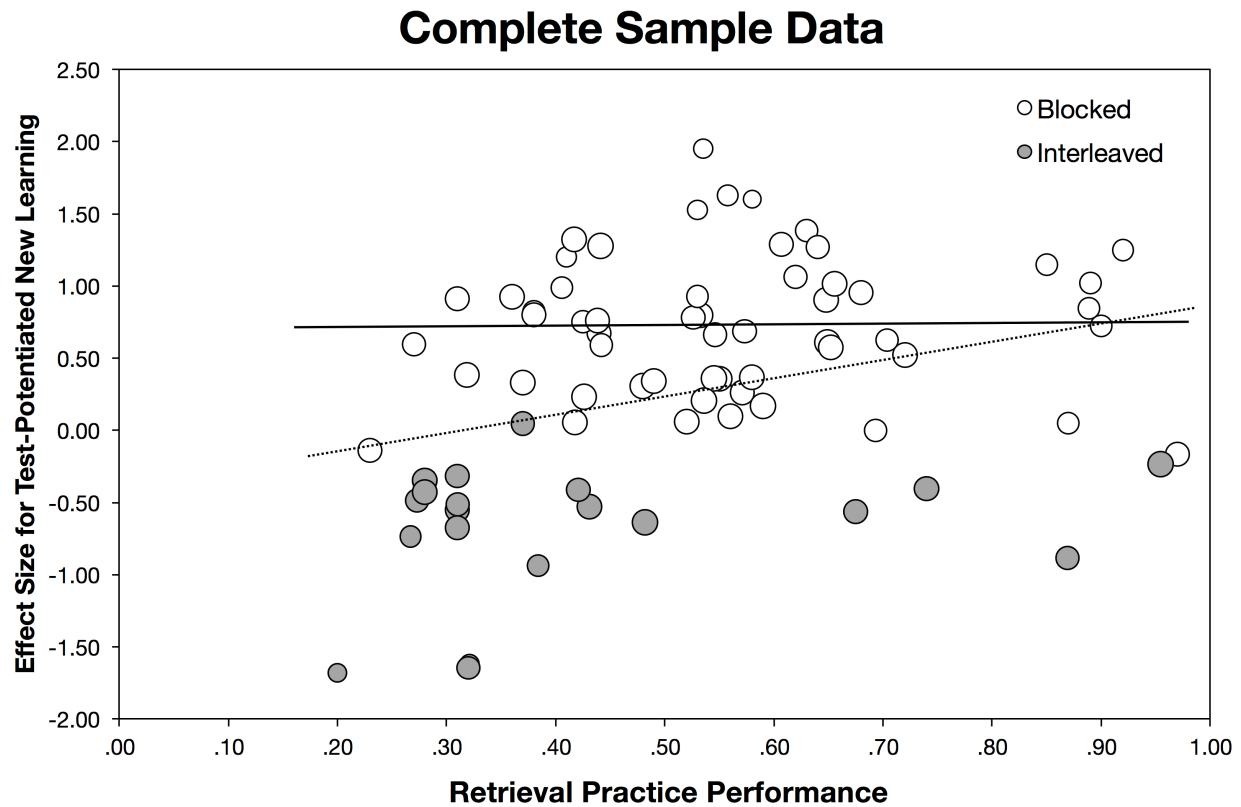
pretest constitutes new learning (and not relearning).



**Figure 3.** A pirate plot displaying effect sizes as a function of study samples. Each data point represents the effect size of an individual study. Jitter was introduced to displace the data points horizontally to improve their visibility. The horizontal line within each set of data shows the mean effect size, and error bars display .95 *CI*.

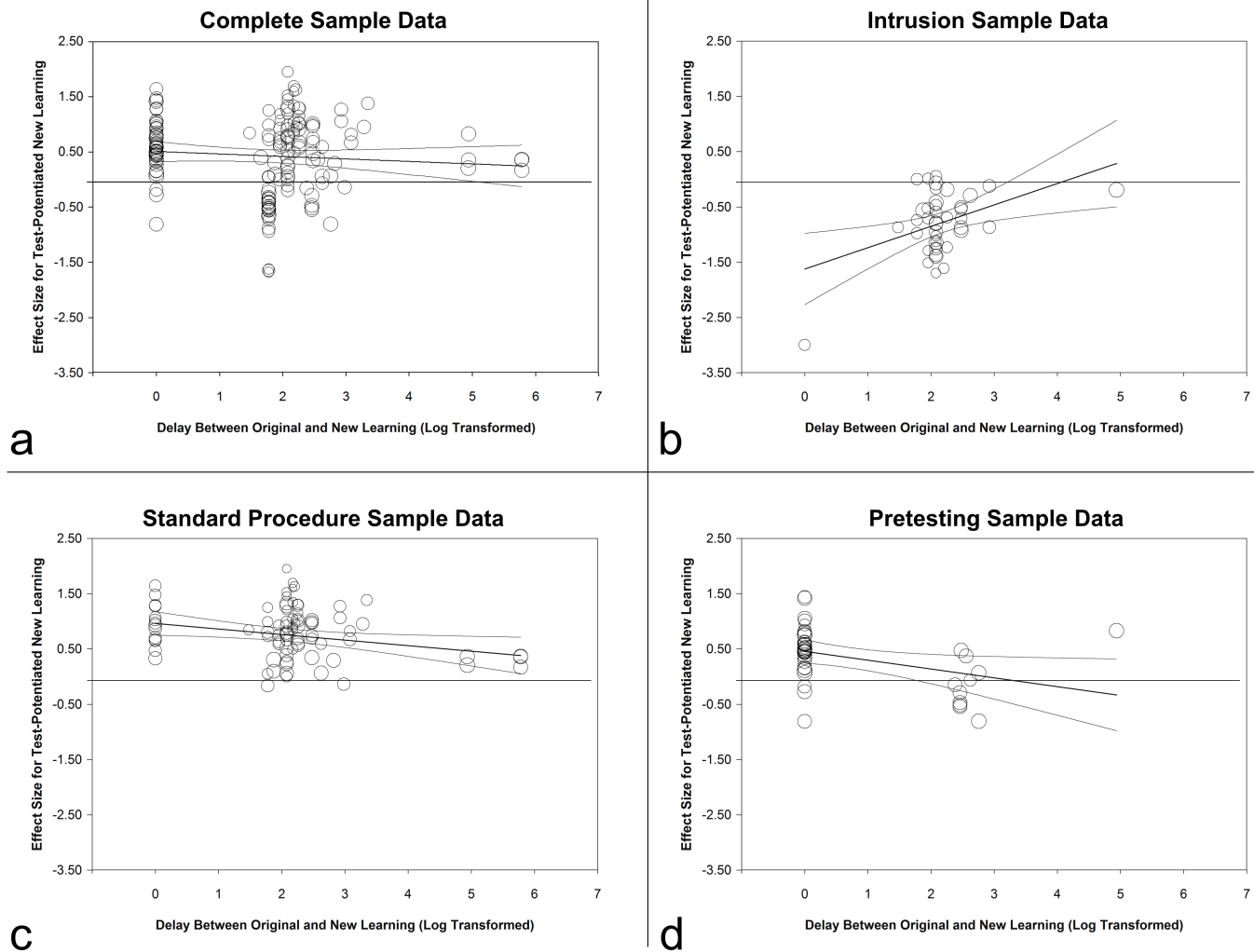


**Figure 4.** Effect of memory load on test-potentiated new learning. Figure 3a shows correct recall data based on the complete sample. Figure 3b shows intrusions data. Lighter lines indicate .95 *CI*.





**Figure 5.** Relation between retrieval practice performance and test-potentiated new learning based on correct recall data. When the entire set of studies were considered together, there was a positive association between the two variables as indicated by the dotted regression line ( $k = 78$ ). However, this association appeared to be driven primarily by studies that used the interleaving design, which are shown in gray. When these studies are omitted from the analysis, the positive association disappeared, as shown by the solid regression line ( $k = 56$ ).





**Figure 6.** Effect of delay between original and new learning on test-potentiated new learning across the four samples. Because delay in seconds was logarithmically transformed, numbers greater than 4.93 on the horizontal axis indicate delays greater than 24 hours. Figure 5a shows data from the complete sample. Figure 5b shows data from the intrusions sample. Figure 5c shows data from the standard procedure sample. Figure 5d shows data from the pretesting sample.

	Interleaving	Research Design	Comparison Task	Relation Original-New Learning	Initial Test Format	Test Format Match	Memory Load	Retrieval Practice Performance	Delay
Resource theories									X
Metacognitive theories						X		X	
Context theories			X	X					X
Integration theories								X	

**Figure 7.** Empirical support for each theory based on results from the moderator analyses. A filled circle indicates strong support (which indicates that results from all samples support the prediction), a half-filled circle indicates limited support (which indicates that some subsamples failed to support the prediction), an X indicates no support, and empty cells indicate that either the theory does not make a specific prediction for the moderator, or that the theory lacks the necessary precision to make a clear prediction.

### Appendix

Studies included in the meta-analysis and their descriptive statistics. Asterisks next to the effect size indicate that the effect size calculation required imputing variability data.

Study	Year	Outcome	Within-Subjects?	Tested <i>N</i>	Control <i>N</i>	Effect Size ( <i>g</i> )	Resource Theory Score	Metacognitive Theory Scores	Context Theory Score	Integration Theory Score
Allen & Arbak	1976	Correct	N	10	10	0.96	2.30	3.00	2.00	3.00
Arkes & Lyons Exp 1										
- MMFR Warning	1979	Correct	N	14	14	1.05*	2.30	3.00	2.00	3.00
Arkes & Lyons Exp 1										
- No MMFR Warning	1979	Correct	N	14	14	0.50*	2.30	3.00	2.00	3.00
Arkes & Lyons Exp 2										
- Mediation	1979	Correct	N	14	14	1.69*	2.30	3.00	2.00	3.00
Arkes & Lyons Exp 2										
- No mediation	1979	Correct	N	14	14	0.90*	2.30	3.00	2.00	3.00
Arkes & Lyons Exp 3										
- Mediation	1979	Correct	N	14	14	1.34*	2.30	3.00	2.00	3.00
Arkes & Lyons Exp 3										
- No mediation	1979	Correct	N	14	14	0.85*	2.30	3.00	2.00	3.00

Study	Year	Outcome	Within-Subjects?	Tested <i>N</i>	Control <i>N</i>	Effect Size ( <i>g</i> )	Resource Theory Score	Metacognitive Theory Scores	Context Theory Score	Integration Theory Score
Aslan & Bauml - 6 yr olds	2015	Correct	N	16	16	0.05	2.60	3.13	3.00	2.87
Aslan & Bauml - 8 yr olds	2015	Correct	N	16	16	0.73	2.60	3.10	3.00	2.90
Aslan & Bauml - Adults	2015	Correct	N	16	16	1.25	2.60	3.08	3.00	2.92
Bauml & Kliegl Exp 2	2013	Correct	Y	30	30	0.52	2.48	2.28	3.00	2.72
Bettencourt Delaney & Chang (unpublished) Exp 1	2015	Correct	N	41	40	-0.56	1.30	3.33	2.00	1.68
Bettencourt Delaney & Chang (unpublished) Exp 2	2015	Correct	Y	88	88	-0.23	1.30	1.05	2.00	1.95
Davis & Chan (unpublished) Exp 1	2015a	Correct	N	53	60	0.91	2.30	4.35	3.00	3.65
Davis & Chan (unpublished) Exp 2	2015a	Correct	N	60	60	0.80	2.30	4.47	3.00	3.53

Study	Year	Outcome	Within-Subjects?	Tested <i>N</i>	Control <i>N</i>	Effect Size ( <i>g</i> )	Resource Theory Score	Metacognitive Theory Scores	Context Theory Score	Integration Theory Score
Davis & Chan										
(unpublished) Exp 3	2015a	Correct	N	48	42	0.26	2.30	4.43	3.00	3.57
LaPaglia & Chan										
(unpublished) Exp 1	2013	Correct	N	36	34	1.29	2.30	4.39	3.00	3.61
LaPaglia & Chan										
(unpublished) Exp 2	2013	Correct	N	64	64	0.61	2.30	4.35	3.00	3.65
LaPaglia & Chan										
(unpublished) Exp 3	2013	Correct	N	54	54	0.57	2.30	4.35	3.00	3.65
LaPaglia & Chan										
(unpublished) Exp 4	2013	Correct	N	69	62	1.02	2.30	3.34	3.00	3.66
LaPaglia & Chan										
(unpublished) Exp 5	2013	Correct	N	29	34	0.69	2.30	4.43	3.00	3.57
LaPaglia & Chan										
(unpublished) Exp 6	2013	Correct	N	38	23	0.66	2.30	4.45	3.00	3.55
Chan & LaPaglia Exp										
4 - 1 test	2011	Correct	Y	38	38	0.17	3.30	3.41	2.00	4.59

Study	Year	Outcome	Within-Subjects?	Tested <i>N</i>	Control <i>N</i>	Effect Size ( <i>g</i> )	Resource Theory Score	Metacognitive Theory Scores	Context Theory Score	Integration Theory Score
Chan & LaPaglia Exp 4 - 3 tests	2011	Correct	Y	35	35	0.36	3.30	3.45	2.00	4.55
Chan & LaPaglia Exp 4 - 6 tests	2011	Correct	Y	40	40	0.37	3.30	3.42	2.00	4.58
Chan Thomas & Bulevich Exp 2	2009	Correct	N	24	24	1.38	3.30	3.37	1.00	4.63
Chan Wilford & Hughes Exp 2	2012	Correct	N	40	40	0.67	3.30	3.56	1.00	4.44
Chan Wilford & Hughes Exp 3	2012	Correct	N	20	20	0.82	3.30	3.62	1.00	4.38
Cho Neely Crocco & Vitrano QJEP Exp 1a - New pairs	2016	Correct	N	50	50	0.33	3.30	4.64	4.00	2.36
Cho Neely Crocco & Vitrano QJEP Exp 2 - New pairs	2016	Correct	N	50	50	0.93	3.30	4.63	4.00	2.37

Study	Year	Outcome	Within- Subjects?	Tested <i>N</i>	Control <i>N</i>	Effect Size ( <i>g</i> )	Resource Theory Score	Metacognitive Theory Scores	Context Theory Score	Integration Theory Score
Davis & Chan Exp 1 - Name priority	2015b	Correct	N	7	17	-1.62	1.30	2.68	2.00	1.32
Davis & Chan Exp 1 - No Priority	2015b	Correct	N	5	17	-1.68	1.30	2.80	2.00	1.20
Davis & Chan Exp 1 - Profession Priority	2015b	Correct	N	15	13	-0.74	1.30	3.73	2.00	1.27
Davis & Chan Exp 2	2015b	Correct	N	30	30	-0.49	1.30	3.73	2.00	1.27
Davis & Chan Exp 3	2015b	Correct	Y	12	12	-0.94	1.30	2.62	2.00	1.38
Davis & Chan Exp 4	2015b	Correct	Y	32	32	-0.53	1.30	2.57	2.00	1.43
Davis & Chan Exp 5	2015b	Correct	N	29	27	-0.88	1.30	3.13	2.00	1.87
Davis & Chan Exp 6	2015b	Correct	Y	50	50	-0.64	1.30	2.52	2.00	1.48
Davis & Chan Exp 1	2015	Correct	N	37	38	-0.55	1.30	3.69	2.00	1.31
Davis & Chan Exp 2	2015	Correct	N	24	27	-1.64	1.30	3.68	2.00	1.32
Davis & Chan Exp 3	2015	Correct	N	31	31	-0.32	1.30	3.69	2.00	1.31
Davis & Chan Exp 4	2015	Correct	N	32	32	0.80	2.30	4.62	3.00	2.38

Study	Year	Outcome	Within-Subjects?	Tested <i>N</i>	Control <i>N</i>	Effect Size ( <i>g</i> )	Resource Theory Score	Metacognitive Theory Scores	Context Theory Score	Integration Theory Score
Davis & Chan Exp 5 -										
Name priority	2015	Correct	N	30	30	-0.41	1.30	3.58	2.00	1.42
Davis & Chan Exp 5 -										
Profession Priority	2015	Correct	N	30	30	0.05	1.30	3.63	2.00	1.37
Divis & Benjamin										
Exp 1	2014	Correct	N	18	21	0.69	1.70	1.00	3.00	2.00
Divis & Benjamin										
Exp 2 - Recognition	2014	Correct	N	7	7	1.19	1.60	1.00	3.00	2.00
Divis & Benjamin										
Exp 2 - Recall	2014	Correct	N	7	7	1.08	1.60	1.00	3.00	2.00
Divis & Benjamin										
Exp 3	2014	Correct	N	25	26	0.00	1.70	1.00	3.00	2.00
Finn & Roediger Exp										
1a	2013	Correct	N	48	62	-0.35	1.30	3.72	2.00	1.28
Finn & Roediger Exp										
1b	2013	Correct	N	17	13	-0.51	1.30	2.69	2.00	1.31



Study	Year	Outcome	Within-Subjects?	Tested <i>N</i>	Control <i>N</i>	Effect Size ( <i>g</i> )	Resource Theory Score	Metacognitive Theory Scores	Context Theory Score	Integration Theory Score
Finn & Roediger Exp										
2	2013	Correct	N	34	43	-0.34	1.30	2.72	2.00	1.28
Finn & Roediger Exp										
3	2013	Correct	N	26	25	-0.51	1.30	2.69	2.00	1.31
Finn & Roediger Exp										
4	2013	Correct	N	53	58	-0.43	2.30	2.72	1.00	2.28
Finn & Roediger Exp										
5	2013	Correct	N	42	44	-0.68	2.30	3.69	1.00	2.31
Finn & Roediger Exp										
6	2013	Correct	N	41	43	-0.40	1.30	3.26	2.00	1.74
Gordon & Thomas										
Exp 3	2014	Correct	N	42	38	0.95	3.30	4.32	1.00	4.68
Gordon unpublished										
dissertation Exp 2	2016	Correct	N	29	29	1.06	3.30	3.38	2.00	4.62
Gordon unpublished										
dissertation Exp 4	2016	Correct	N	27	33	1.27	3.30	4.36	2.00	4.64

Study	Year	Outcome	Within-Subjects?	Tested <i>N</i>	Control <i>N</i>	Effect Size ( <i>g</i> )	Resource Theory Score	Metacognitive Theory Scores	Context Theory Score	Integration Theory Score
Grimaldi & Karpicke										
Exp 1	2012	Correct	N	16	16	0.94	2.00	3.00	2.00	4.00
Grimaldi & Karpicke										
Exp 2	2012	Correct	N	30	30	-0.18	1.00	3.00	3.00	3.00
Grimaldi & Karpicke										
Exp 3	2012	Correct	N	18	18	-0.07	2.00	3.00	2.00	3.00
Hays Kornell & Bjork										
Exp 1	2013	Correct	Y	70	70	-0.81	1.00	2.00	1.00	3.00
Hays Kornell & Bjork										
Exp 2	2013	Correct	Y	45	45	0.07	1.00	2.00	1.00	4.00
Huff Davis & Meade										
Exp 3	2013	Correct	N	36	36	-0.14*	3.30	3.77	2.00	4.23
Knight Ball Brewer										
DeWitt & Marsh Exp										
1a - Related	2012	Correct	Y	30	30	0.47	2.00	2.00	2.00	3.00

Study	Year	Outcome	Within-Subjects?	Tested <i>N</i>	Control <i>N</i>	Effect Size ( <i>g</i> )	Resource Theory Score	Metacognitive Theory Scores	Context Theory Score	Integration Theory Score
Knight Ball Brewer										
DeWitt & Marsh Exp										
1a - Unrelated	2012	Correct	Y	30	30	-0.81	1.00	2.00	3.00	2.00
Knight Ball Brewer										
DeWitt & Marsh Exp										
1b - Related	2012	Correct	Y	35	35	0.81	2.00	2.00	2.00	4.00
Knight Ball Brewer										
DeWitt & Marsh Exp										
1b - Unrelated	2012	Correct	Y	35	35	-0.27	1.00	2.00	3.00	3.00
Kornell Exp 1	2014	Correct	Y	23	23	-0.15	2.00	2.00	2.00	3.00
Kornell Exp 2	2014	Correct	Y	31	31	0.37	2.00	2.00	2.00	3.00
Kornell Exp 3a	2014	Correct	Y	76	76	0.83	2.00	2.00	2.00	3.00
Kornell Exp 3b	2014	Correct	Y	52	52	0.47	2.00	2.00	2.00	3.00
Kornell Hays & Bjork										
Exp 1	2009	Correct	Y	25	25	0.50	1.00	2.00	3.00	3.00

Study	Year	Outcome	Within-Subjects?	Tested <i>N</i>	Control <i>N</i>	Effect Size ( <i>g</i> )	Resource Theory Score	Metacognitive Theory Scores	Context Theory Score	Integration Theory Score
Kornell Hays & Bjork										
Exp 2	2009	Correct	Y	20	20	0.15	2.00	2.00	2.00	3.00
Kornell Hays & Bjork										
Exp 3	2009	Correct	Y	15	15	1.02	2.00	2.00	2.00	4.00
Kornell Hays & Bjork										
Exp 4	2009	Correct	Y	15	15	0.53	2.00	2.00	2.00	3.00
Kornell Hays & Bjork										
Exp 5	2009	Correct	Y	30	30	0.59	2.00	2.00	2.00	3.00
Kornell Hays & Bjork										
Exp 6	2009	Correct	N	42	42	0.47	2.00	3.00	2.00	3.00
Lane Mather Villa &										
Morita Exp 2	2001	Correct	N	72	72	0.29	3.30	3.00	1.00	4.00
Lehman Smith &										
Karpicke	2014	Correct	N	36	36	0.91	3.70	3.69	2.00	4.31
Nunes & Weinstein										
Exp 1	2012	Correct	N	31	31	0.78	3.70	3.47	2.00	4.53

Study	Year	Outcome	Within-Subjects?	Tested <i>N</i>	Control <i>N</i>	Effect Size ( <i>g</i> )	Resource Theory Score	Metacognitive Theory Scores	Context Theory Score	Integration Theory Score
Nunes & Weinstein										
Exp 2	2012	Correct	N	19	16	0.00	3.70	3.31	2.00	4.69
Nunes & Weinstein										
Exp 3	2012	Correct	N	15	15	0.84	3.70	3.11	2.00	4.89
Pashler Kang &										
Mozer Exp 1	2013	Correct	Y	56	56	0.36	4.08	2.45	2.00	4.55
Pashler Kang &										
Mozer Exp 2	2013	Correct	Y	69	69	0.21	4.08	3.46	2.00	4.54
Pastotter & Bauml										
(unpublished)	2016	Correct	Y	90	90	0.31	2.70	3.52	3.00	3.48
Pastotter Bauml &										
Hanslmayr	2008	Correct	Y	48	48	0.40	1.30	0.00	2.00	2.00
Pastotter Schicker										
Niedernhuber &										
Bauml	2011	Correct	N	18	18	0.93	2.70	4.47	3.00	2.53
Pastotter Weber &										
Bauml - Healthy	2013	Correct	Y	12	12	0.63	2.48	3.30	2.00	3.70

Study	Year	Outcome	Within-Subjects?	Tested <i>N</i>	Control <i>N</i>	Effect Size ( <i>g</i> )	Resource Theory Score	Metacognitive Theory Scores	Context Theory Score	Integration Theory Score
Pastotter Weber & Bauml - High Brain Damaged	2013	Correct	Y	12	12	0.59	2.48	3.56	2.00	3.44
Pastotter Weber & Bauml - Low Brain Damaged	2013	Correct	Y	12	12	0.75	2.48	3.58	2.00	3.43
Pierce & Hawthorne Exp 1	2016	Correct	N	46	46	0.39	2.70	4.68	2.00	3.32
Pierce & Hawthorne Exp 2	2016	Correct	N	46	46	0.76	3.70	4.56	1.00	4.44
Potts & Shanks Exp 1	2012	Correct	Y	24	24	0.44	1.00	2.00	3.00	2.00
Potts & Shanks Exp 2a	2012	Correct	Y	30	30	0.38	1.00	2.00	3.00	2.00
Potts & Shanks Exp 2b	2012	Correct	Y	24	24	0.27	1.00	2.00	3.00	2.00
Potts & Shanks Exp 3 - Exp Paced	2012	Correct	Y	16	16	0.45	1.00	2.00	3.00	2.00

Study	Year	Outcome	Within-Subjects?	Tested <i>N</i>	Control <i>N</i>	Effect Size ( <i>g</i> )	Resource Theory Score	Metacognitive Theory Scores	Context Theory Score	Integration Theory Score
Potts & Shanks Exp 3										
- Self Paced	2012	Correct	Y	24	24	0.58	1.00	2.00	3.00	2.00
Potts unpublished dissertation Exp 4	2013	Correct	Y	30	30	0.34	1.00	2.00	3.00	2.00
Potts unpublished dissertation Exp 6	2013	Correct	Y	130	130	0.11	1.00	2.00	3.00	2.00
Potts unpublished dissertation Exp 7	2013	Correct	Y	24	24	0.16	1.00	2.00	3.00	2.00
Potts unpublished dissertation Exp 8	2013	Correct	Y	118	118	0.06	1.00	2.00	3.00	2.00
Richland Kornell & Kao Exp 1	2009	Correct	N	36	27	1.44	3.00	3.00	3.00	2.00
Richland Kornell & Kao Exp 2	2009	Correct	N	33	26	0.63	3.00	3.00	3.00	2.00
Richland Kornell & Kao Exp 3	2009	Correct	N	31	33	0.80	3.00	3.00	3.00	2.00

Study	Year	Outcome	Within-Subjects?	Tested <i>N</i>	Control <i>N</i>	Effect Size ( <i>g</i> )	Resource Theory Score	Metacognitive Theory Scores	Context Theory Score	Integration Theory Score
Richland Kornell & Kao Exp 4	2009	Correct	N	79	79	0.44	3.00	3.00	3.00	2.00
Richland Kornell & Kao Exp 5	2009	Correct	N	24	26	0.59	3.00	3.00	3.00	3.00
Robbins & Irvin	1976	Correct	N	16	16	0.99*	2.30	3.59	3.00	3.41
Szpunar Jing & Schacter	2014	Correct	N	18	18	1.30	3.60	3.00	2.00	4.00
Szpunar Khan & Schacter Exp 1	2013	Correct	N	16	16	1.15	3.60	3.15	2.00	4.85
Szpunar Khan & Schacter Exp 2	2013	Correct	N	16	16	1.02	3.60	3.11	2.00	4.89
Szpunar McDermott & Roediger Exp 1a	2008	Correct	N	12	12	1.53	3.70	3.47	2.00	4.53
Szpunar McDermott & Roediger Exp 1b	2008	Correct	N	12	12	1.20	2.70	3.59	3.00	3.41
Szpunar McDermott & Roediger Exp 2	2008	Correct	N	12	12	1.95	3.70	3.47	2.00	4.54



Study	Year	Outcome	Within-Subjects?	Tested <i>N</i>	Control <i>N</i>	Effect Size ( <i>g</i> )	Resource Theory Score	Metacognitive Theory Scores	Context Theory Score	Integration Theory Score
Szpunar McDermott										
& Roediger Exp 3	2008	Correct	N	12	12	1.45	3.70	3.00	2.00	3.00
Tulving & Watkins										
Exp 1	1974	Correct	N	8	8	1.60*	2.30	3.42	3.00	3.58
Tulving & Watkins										
Exp 2	1974	Correct	Y	32	32	0.10*	2.30	2.44	3.00	3.56
Vaughn & Rawson										
Exp 1 - 1 pretrial	2012	Correct	Y	26	26	-0.55	2.00	2.00	2.00	2.00
Vaughn & Rawson										
Exp 1 - 3 pretrials	2012	Correct	Y	26	26	-0.29	2.00	2.00	2.00	2.00
Vaughn & Rawson										
Exp 2 - Delayed study	2012	Correct	Y	34	34	-0.51	2.00	2.00	2.00	2.00
Vaughn & Rawson										
Exp 2 - Immediate study	2012	Correct	Y	32	32	0.57	3.00	2.00	3.00	2.00
Vaughn & Rawson										
Exp 3 - Delayed study	2012	Correct	Y	30	30	-0.47	2.00	2.00	2.00	2.00

Study	Year	Outcome	Within-Subjects?	Tested <i>N</i>	Control <i>N</i>	Effect Size ( <i>g</i> )	Resource Theory Score	Metacognitive Theory Scores	Context Theory Score	Integration Theory Score
Vaughn & Rawson										
Exp 3 - immediate study	2012	Correct	Y	29	29	0.75	3.00	2.00	3.00	2.00
Vaughn Hausman & Kornell Exp 1a	2016	Correct	Y	47	47	0.61	2.00	2.00	2.00	3.00
Vaughn Hausman & Kornell Exp 1b	2016	Correct	Y	49	49	0.50	2.00	2.00	2.00	3.00
Vaughn Hausman & Kornell Exp 2a	2016	Correct	Y	97	97	1.42	2.00	2.00	2.00	3.00
Vaughn Hausman & Kornell Exp 2b	2016	Correct	Y	68	68	0.75	2.00	2.00	2.00	3.00
Vaughn Hausman & Kornell Exp 3	2016	Correct	Y	57	57	0.45	2.00	2.00	2.00	3.00
Vaughn Hausman & Kornell Exp 4	2016	Correct	Y	43	43	1.05	2.00	2.00	2.00	3.00
Wahlheim & Jacoby										
Exp 1	2010	Correct	N	36	36	-0.16	2.30	3.03	3.00	3.97

Study	Year	Outcome	Within-Subjects?	Tested <i>N</i>	Control <i>N</i>	Effect Size ( <i>g</i> )	Resource Theory Score	Metacognitive Theory Scores	Context Theory Score	Integration Theory Score
Wahlheim Exp 1	2015	Correct	Y	48	48	0.34	2.30	3.51	3.00	2.49
Wahlheim Exp 2	2015	Correct	Y	36	36	0.06	2.30	3.48	3.00	2.52
Weinstein Gilmore Szpunar & McDermott Exp 1 - Not Warned	2014	Correct	N	60	65	1.32	3.70	4.58	2.00	4.42
Weinstein Gilmore Szpunar & McDermott Exp 1 - Warned	2014	Correct	N	63	62	0.23	3.70	4.57	2.00	4.43
Weinstein Gilmore Szpunar & McDermott Exp 2 - Not Warned	2014	Correct	N	85	81	1.28	3.70	4.56	2.00	4.44
Weinstein Gilmore Szpunar &	2014	Correct	N	71	88	0.06	3.70	4.58	2.00	4.42

Study	Year	Outcome	Within-Subjects?	Tested <i>N</i>	Control <i>N</i>	Effect Size ( <i>g</i> )	Resource Theory Score	Metacognitive Theory Scores	Context Theory Score	Integration Theory Score
McDermott Exp 2 - Warned Weinstein McDermott & Szpunar	2011	Correct	N	16	16	1.63	2.60	4.44	3.00	3.56
Weinstein McDermott Szpunar Bauml & Pastotter Exp 1	2015	Correct	N	17	18	0.59	2.70	1.00	2.00	3.00
Weinstein McDermott Szpunar Bauml & Pastotter Exp 2	2015	Correct	N	20	20	0.13	2.70	1.00	2.00	3.00
Weinstein McDermott Szpunar Bauml & Pastotter Exp 3	2015	Correct	N	20	20	-0.10	1.70	1.00	3.00	2.00
Wissman & Rawson Exp 1	2015b	Correct	N	28	25	1.02	4.60	3.00	3.00	4.00

Study	Year	Outcome	Within-Subjects?	Tested <i>N</i>	Control <i>N</i>	Effect Size ( <i>g</i> )	Resource Theory Score	Metacognitive Theory Scores	Context Theory Score	Integration Theory Score
Wissman & Rawson										
Exp 2 - Section vs. Whole	2015b	Correct	N	29	28	0.97	4.60	3.00	3.00	4.00
Wissman & Rawson										
Exp 2 - Section-Math vs. Whole-Math	2015b	Correct	N	30	28	0.98	3.60	3.00	2.00	4.00
Wissman & Rawson										
Exp 3	2015b	Correct	N	30	29	1.29	4.60	3.00	3.00	4.00
Wissman & Rawson										
Exp 3 Section-delay vs. Whole-delay	2015b	Correct	N	28	27	1.48	4.60	3.00	3.00	4.00
Wissman & Rawson										
Exp 4	2015b	Correct	N	31	31	1.64	4.60	3.00	3.00	4.00
Wissman & Rawson										
Exp 5	2015b	Correct	N	27	27	0.66	4.60	3.00	3.00	4.00
Wissman & Rawson										
Exp 6	2015b	Correct	N	32	30	0.70	4.60	3.00	3.00	4.00

Study	Year	Outcome	Within-Subjects?	Tested <i>N</i>	Control <i>N</i>	Effect Size ( <i>g</i> )	Resource Theory Score	Metacognitive Theory Scores	Context Theory Score	Integration Theory Score
Wissman & Rawson unpublished Exp 2 - Section vs. Whole	2015a	Correct	N	31	32	1.28	4.48	3.00	3.00	4.00
Wissman & Rawson unpublished Exp 2 - Section-Pre vs. Whole-Pre	2015a	Correct	N	30	31	0.47	4.48	4.00	3.00	4.00
Wissman & Rawson unpublished Exp 3 - Section-Delay vs. Whole	2015a	Correct	N	37	29	0.65	4.60	4.00	3.00	4.00
Wissman & Rawson unpublished Exp 1 - Section-Section vs. Whole-Section	2015a	Correct	N	33	32	0.86	4.60	4.00	3.00	4.00
Wissman & Rawson unpublished Exp 1 -	2015a	Correct	N	30	32	1.07	4.60	3.00	3.00	4.00

Study	Year	Outcome	Within-Subjects?	Tested <i>N</i>	Control <i>N</i>	Effect Size ( <i>g</i> )	Resource Theory Score	Metacognitive Theory Scores	Context Theory Score	Integration Theory Score
Section-Whole vs. Whole-Whole										
Wissman Rawson & Pyc Exp 1a	2011	Correct	N	64	65	0.98	2.48	4.00	3.00	3.00
Wissman Rawson & Pyc Exp 1b	2011	Correct	N	21	19	0.69	3.48	4.00	2.00	4.00
Wissman Rawson & Pyc Exp 2	2011	Correct	N	54	59	1.02	3.60	4.00	2.00	4.00
Wissman Rawson & Pyc Exp 3	2011	Correct	N	29	30	0.75	3.48	4.00	2.00	4.00
Wissman Rawson & Pyc Exp 4	2011	Correct	N	23	26	0.97	3.48	4.00	2.00	4.00
Yue Soderstrom & Bjork Exp 2	2015	Correct	N	23	24	0.59	2.30	3.73	3.00	2.27
Aslan & Bauml - 6 yr olds	2016	Intrusions	N	16	16	0.00				

Study	Year	Outcome	Within-Subjects?	Tested <i>N</i>	Control <i>N</i>	Effect Size ( <i>g</i> )	Resource Theory Score	Metacognitive Theory Scores	Context Theory Score	Integration Theory Score
Aslan & Bauml - 8 yr olds	2016	Intrusions	N	16	16	-0.74				
Aslan & Bauml - Adults	2016	Intrusions	N	16	16	-0.98				
Bauml & Kliegl Exp 2	2013	Intrusions	Y	30	30	-3.00*				
LaPaglia & Chan (unpublished) Exp 2	2013	Intrusions	N	64	64	-0.07				
Divis & Benjamin Exp 1	2014	Intrusions	N	18	21	-0.18				
Gordon unpublished dissertation Exp 2	2016	Intrusions	N	29	29	-0.53				
Gordon unpublished dissertation Exp 4	2016	Intrusions	N	27	33	-0.86				
Lehman Smith & Karpicke	2014	Intrusions	N	36	36	-0.12				
Nunes & Weinstein Exp 1	2012	Intrusions	N	31	31	-1.25				



Study	Year	Outcome	Within-Subjects?	Tested <i>N</i>	Control <i>N</i>	Effect Size ( <i>g</i> )	Resource Theory Score	Metacognitive Theory Scores	Context Theory Score	Integration Theory Score
Nunes & Weinstein										
Exp 2	2012	Intrusions	N	19	16	-1.38				
Nunes & Weinstein										
Exp 3	2012	Intrusions	N	15	15	-0.47				
Pashler Kang &										
Mozer Exp 1	2013	Intrusions	Y	56	56	-0.87				
Pastotter & Bauml										
(unpublished)	2016	Intrusions	Y	90	90	-0.71				
Pastotter Schicker										
Niedernhuber &										
Bauml	2011	Intrusions	N	18	18	-0.87				
Pastotter Weber &										
Bauml - Healthy	2013	Intrusions	N	12	12	-0.19				
Pastotter Weber &										
Bauml - High Brain										
Damaged	2013	Intrusions	N	12	12	-0.56				

Study	Year	Outcome	Within-Subjects?	Tested <i>N</i>	Control <i>N</i>	Effect Size ( <i>g</i> )	Resource Theory Score	Metacognitive Theory Scores	Context Theory Score	Integration Theory Score
Pastotter Weber & Bauml - Low Brain Damaged	2013	Intrusions	N	12	12	-0.70				
Pierce & Hawthorne Exp 1	2016	Intrusions	N	46	46	-1.29				
Pierce & Hawthorne Exp 2	2016	Intrusions	N	46	46	0.01				
Szpunar Khan & Schacter Exp 1	2013	Intrusions	N	16	16	-1.51				
Szpunar Khan & Schacter Exp 2	2013	Intrusions	N	16	16	-0.96				
Szpunar McDermott & Roediger Exp 1a	2008	Intrusions	N	12	12	-0.81				
Szpunar McDermott & Roediger Exp 1b	2008	Intrusions	N	12	12	-0.69				
Szpunar McDermott & Roediger Exp 2	2008	Intrusions	N	12	12	-1.23				

Study	Year	Outcome	Within-Subjects?	Tested <i>N</i>	Control <i>N</i>	Effect Size ( <i>g</i> )	Resource Theory Score	Metacognitive Theory Scores	Context Theory Score	Integration Theory Score
Szpunar McDermott & Roediger Exp 3	2008	Intrusions	N	12	12	-1.24				
Wahlheim Exp 1	2015	Intrusions	Y	48	48	-0.50				
Wahlheim Exp 2	2015	Intrusions	Y	36	36	-0.54				
Weinstein Gilmore Szpunar & McDermott Exp 1 - Not Warned	2014	Intrusions	N	60	65	-1.13				
Weinstein Gilmore Szpunar & McDermott Exp 1 - Warned	2014	Intrusions	N	63	62	-1.02				
Weinstein Gilmore Szpunar & McDermott Exp 2 - Not Warned	2014	Intrusions	N	85	81	-1.70				

Study	Year	Outcome	Within-Subjects?	Tested <i>N</i>	Control <i>N</i>	Effect Size ( <i>g</i> )	Resource Theory Score	Metacognitive Theory Scores	Context Theory Score	Integration Theory Score
Weinstein Gilmore										
Szpunar & McDermott Exp 2 -										
Warned	2014	Intrusions	N	71	88	-0.93				
Weinstein McDermott & Szpunar	2011	Intrusions	N	16	16	-0.29				
Weinstein McDermott Szpunar Bauml & Pastotter Exp 1	2015	Intrusions	N	17	18	-1.39				
Weinstein McDermott Szpunar Bauml & Pastotter Exp 2	2015	Intrusions	N	20	20	-0.79				
Weinstein McDermott Szpunar Bauml & Pastotter Exp 3	2015	Intrusions	N	20	20	-1.14				
Wissman & Rawson Exp 4	2015b	Intrusions	N	31	31	-0.42*				

Study	Year	Outcome	Within-Subjects?	Tested <i>N</i>	Control <i>N</i>	Effect Size ( <i>g</i> )	Resource Theory Score	Metacognitive Theory Scores	Context Theory Score	Integration Theory Score
Wissman Rawson & Pyc Exp 1a	2011	Intrusions	N	64	65	-1.61*				
Wissman Rawson & Pyc Exp 1b	2011	Intrusions	N	21	19	0.05*				
Wissman Rawson & Pyc Exp 3	2011	Intrusions	N	29	30	-0.59*				
Wissman Rawson & Pyc Exp 4	2011	Intrusions	N	23	26	-0.16*				

---

