The Effect of Testing Versus Restudy on Retention: A Meta-Analytic Review of the Testing Effect

Article *in* Psychological Bulletin · August 2014

DOI: 10.1037/a0037559 · Source: PubMed

Christopher A Rowland
Eckerd College

The Effect of Testing Versus Restudy on Retention: A Meta-Analytic Review of the Testing Effect

Christopher A. Rowland

Online First Publication, August 25, 2014. http://dx.doi.org/10.1037/a0037559

Psychological Bulletin

0033-2909/14/$12.00 http://dx.doi.org/10.1037/a0037559

The Effect of Testing Versus Restudy on Retention: A Meta-Analytic Review of the Testing Effect

Christopher A. Rowland
Colorado State University

Engaging in a test over previously studied information can serve as a potent learning event, a phenomenon referred to as the *testing effect*. Despite a surge of research in the past decade, existing theories have not yet provided a cohesive account of testing phenomena. The present study uses meta-analysis to examine the effects of testing versus restudy on retention. Key results indicate support for the role of effortful processing as a contributor to the testing effect, with initial recall tests yielding larger testing benefits than recognition tests. Limited support was found for existing theoretical accounts attributing the testing effect to enhanced semantic elaboration, indicating that consideration of alternative mechanisms is warranted in explaining testing effects. Future theoretical accounts of the testing effect may benefit from consideration of episodic and contextually derived contributions to retention resulting from memory retrieval. Additionally, the bifurcation model of the testing effect is considered as a viable framework from which to characterize the patterns of results present across the literature.

Keywords: testing effect, retrieval practice, meta-analysis, memory, retrieval

Memory research has repeatedly demonstrated an interdependence between the processes of encoding new information, storing it over time, and accessing it through retrieval. One clear demonstration is the *testing effect*—the finding that retrieving information from memory can, under many circumstances, strengthen one's memory of the retrieved information (for recent reviews, see Roediger & Butler, 2011; Roediger & Karpicke, 2006a). Although the positive effect of testing memory on retention has been known for some time (e.g., for early investigations of testing effects, see Abbott, 1909; Gates, 1917; Spitzer, 1939), research on the topic has grown considerably over the past decade (see Rawson & Dunlosky, 2011).

The testing effect presents as a robust phenomenon, demonstrated using a wide variety of materials, including single-word lists (e.g., Carpenter & DeLosh, 2006; Rowland & DeLosh, 2014b; Rowland, Littrell-Baez, Sensenig, & DeLosh, 2014; Zaromb & Roediger, 2010), paired associates (e.g., Allen, Mahler, & Estes, 1969; Carpenter, 2009; Carpenter, Pashler, & Vul, 2006; Carrier & Pashler, 1992; Pyc & Rawson, 2010; Toppino & Cohen, 2009), prose passages (e.g., Glover, 1989; Roediger & Karpicke, 2006b), and nonverbal materials (e.g., Carpenter & Pashler, 2007; Kang, 2010). Despite this variability in materials used, and in many cases the experimental procedures employed, most studies on the testing effect can be described as having a few discrete phases. First, information of some type is presented to participants for an initial learning opportunity. At some time following initial learning, an intervening phase occurs during which the information can either be re-presented for additional study (restudy condition), subjected to a memory test (test condition; with or without corrective feedback), or not reexposed at all (no test, or study-only condition). Later, following a retention interval, a final memory assessment is given for the information previously learned. The testing effect refers to the common finding that information subjected to a test at the intervening phase of an experiment is better remembered on the final assessment compared with information granted restudy, or information not returned to at all after initial learning. Thus, tests can serve as effective learning events in and of themselves.

Despite the generality of the testing effect, much of the existing research on the effect has not had a clear theoretical orientation (Pyc & Rawson, 2011; Rawson & Dunlosky, 2011). Those investigations that have been conducted with a theoretical focus have presented, in some instances, frameworks that are generally applicable but broadly defined such that testability is limited (e.g., the retrieval hypothesis; see Glover, 1989). Conversely, other theories are explicitly defined but with unclear applicability beyond a specific subset of testing effect investigations (e.g., the well-defined mediator effectiveness hypothesis, applied to verbal paired-associate learning; see Pyc & Rawson, 2010). As such, a cohesive, mechanistic account of testing phenomena remains undeveloped. Despite the relative lack of theoretical emphasis in the testing effect literature, however, much effort has been devoted toward identifying the experimental conditions under which testing may or may not be beneficial for memory (Pyc & Rawson, 2009; Rawson & Dunlosky, 2011; Roediger & Karpicke, 2006a). On the whole, the literature on the testing effect has provided a rich description of conditions and factors that may moderate or mediate testing benefits on retention but has yielded limited development concerning the underlying mechanisms driving the effect.

The present meta-analysis focuses on theoretical issues pertaining to the testing effect. That is, theoretical characterizations of the

I thank Edward DeLosh and Matthew Rhodes for their valuable comments throughout the process of developing this article.

Correspondence concerning this article should be addressed to Christopher A. Rowland, Department of Psychology, Colorado State University, Fort Collins, CO 80521. E-mail: rowlandc@colostate.edu

testing effect are reviewed and assessed, and an additional emphasis is placed on evaluating boundary conditions of test-induced learning that may be especially useful for future theorizing. I begin by discussing the scope of the presented meta-analysis, given that research on memory retrieval encompasses a methodologically diverse literature that is not easily synthesized at a quantitative level. Next, I describe a number of contemporary theoretical characterizations of the testing effect, and highlight some key predictions derived from them. I follow with an outline of three groups of factors that have been identified as theoretically informative boundary conditions of the testing effect: the impact of experimental design, the length of the retention interval, and the influence of stimulus characteristics. In each section, I indicate relevant moderator variables that were examined in the meta-analysis.

Scope of the Present Meta-Analysis

As noted above, the testing effect is commonly studied using a general experimental framework with an initial learning phase, an intervening phase (in which retrieval occurs), and an assessment phase. However, a diverse set of specific methodological procedures have been adapted within this general framework. One difference across studies is with regard to the type of control condition employed. When tested information is contrasted with information that is only presented during initial learning (i.e., a no-test or study-only control), the effect of initial testing becomes confounded with item exposure such that retrieved information in the test condition is reexposed to a participant, whereas nontested information is not. This can potentially produce an upwardly biased estimate of the retention benefit that results from the act of retrieval itself. This shortcoming can be addressed by employing a restudy control condition, though in such cases the bias reverses. That is, a restudy opportunity grants reexposure to all control condition information, whereas an initial test only grants reexposure to successfully retrieved information (unless feedback is provided).

Still other methods used to investigate the testing effect have utilized comparison conditions in which the number, format, spacing, or other aspects of initial tests are varied, such that all experimental conditions receive testing to some extent. These studies often utilize combinations of criterion learning (e.g., Pyc & Rawson, 2009; Vaughn & Rawson, 2011; e.g., “Are two successful tests more effective than one?”), dropout schedules of learning (e.g., Karpicke & Roediger, 2008; Pyc & Rawson, 2007; e.g., “Does additional study or testing after a successful retrieval benefit retention?”), manipulations of the spacing or difficulty of retrieval practice trials (e.g., Carpenter & DeLosh, 2006, Experiment 3; Karpicke & Bauernschmidt, 2011; Pyc & Rawson, 2009), and other factors (note that such designs need not be mutually exclusive). All such methods for studying the effects of testing have notable strengths and are typically utilized to address a specific research question.

Given the possible (but nonexhaustive) types of control conditions and paradigms outlined above, testing effect studies may compare testing to initial study only; testing to equivalent duration of restudy; or certain forms or amounts of testing to other various forms or amounts of testing. Such methodological variability does not allow for quantitative synthesis in a meta-analysis oriented toward addressing theoretical issues derived from specific design characteristics, as is the case in the present investigation. That is, meta-analysis requires that all included studies predict a conceptually common effect (subject to sources of error and moderation, of course, as described in the Method section). Meta-analyses synthesizing overly broad or diverse study protocols provide little value in assessing theoretical explanations for specific effects of interest (Wood & Eagly, 2009). In the present case, variable classes of control conditions were deemed to be estimating sufficiently different effects so as to be unsuitable for addressing the goals of the meta-analysis. In particular, the present meta-analysis is designed in part to assess the extent to which testing is beneficial to retention beyond the effects of reexposure alone. Thus, the corresponding subset of the literature appropriate for informing this question is considered (i.e., studies comparing the effect of testing versus an equivalent duration of restudy on retention). Given that retrieval phases are employed in research protocols extending far beyond the testing effect literature, an additional point of concern arises from the clustering of methodological characteristics and control conditions that are common to various research topics.

Rawson and Dunlosky (2011) list a number of research areas that often utilize retrieval practice in some form, including “retrieval-induced forgetting, generation effects, adjunct questions, hypercorrection effects, underconfidence with practice, and hypermnesia” (p. 284). Such investigations include conditions that often map onto, broadly, certain types of testing effect studies (e.g., retrieval-induced forgetting studies typically employ test and no-test conditions; hypermnesia studies may manipulate the number of tests given). The present investigation takes the position that including overly diverse retrieval methodologies that capture many classes of studies, in a meta-analysis designed to address specific theoretical questions that pertain to and derive from the testing effect literature, would largely impair the utility of the meta-analysis (see Wood & Eagly, 2009). A more inclusive consideration of relevant aspects of studies in variable domains is better addressed by a narrative rather than quantitative review, or in a quantitative review specifically designed to assess broad aspects of memory retrieval. Furthermore, studies in a given research area often cluster together along a number of design characteristics, an issue of concern in any meta-analytic study, including the present one (for further consideration, see the Results and Discussion sections). For example, a question of interest in the testing effect literature pertains to the role of test types, and there is therefore some variability in manipulations of test types across the testing effect literature. However, studies in the related literature of retrieval-induced forgetting almost universally cluster around the specific test type of cued recall, as well as a number of other specific design characteristics, coupled with a no-test control condition (following M. C. Anderson, Bjork, & Bjork’s, 1994, retrieval practice paradigm). An all-inclusive attempt to capture every study utilizing conditions that include retrieval phases in their various instantiations would not be appropriate for present purposes. Instead, the present meta-analysis is intended to capture a portion of the literature on memory retrieval that utilizes a methodological framework suitable for addressing a number of open theoretical questions relating to the testing effect.

Studies included in the present meta-analysis were therefore those utilizing an experimental protocol that was deemed especially common in the testing effect literature, examining the extent to which testing impacts retention beyond the effect of restudy. Studies must (a) include a restudy control condition that equates the duration of the test and restudy opportunities, (b) treat all items within conditions equally, and (c) manipulate testing versus restudy at the time of the intervening phase. Studies utilizing criterion learning, item dropout schedules, or similarly manipulating only the number or format of tests administered thus do not fit into this framework. Such studies may themselves benefit from an independent quantitative review designed to capture the nuances within such areas of research, and to examine related theoretical questions concerning memory retrieval effects. Additionally, classroom studies were not included given a relative lack of control over participant behavior, a factor deemed important given the theoretical orientation of this investigation. For readers interested in applied issues related to testing, Bangert-Drowns, Kulik, and Kulik (1991) report a classroom study meta-analysis of testing, and a number recent educationally oriented publications (Karpicke & Grimaldi, 2012; McDaniel, Roediger, & McDermott, 2007; Rawson & Dunlosky, 2012; Roediger, Agarwal, Kang, & Marsh, 2010; Roediger, Putnam, & Smith, 2011) provide excellent review and coverage of the utility of testing in promoting student learning. Given the aim of the meta-analysis, then, the restrictions placed on study inclusion were designed to capture a circumscribed portion of the literature on retrieval effects. The shortcomings of this approach are noted, where relevant, when interpreting the results of the meta-analysis. It is worth noting up front, however, that the current approach likely underestimates the magnitude of testing advantages, as a result of adopting studies with the more conservative restudy control condition. The observation of positive effects under these conservative conditions firmly establishes the robust character of the testing effect, and one can reasonably expect that the practical application of the procedures investigated may yield substantially larger effects.

Key Theories of the Testing Effect

Many early investigations of the testing effect compared the retention of information that was initially studied but not revisited during learning (no test, or study-only condition) to that which was initially studied then tested, often resulting in a benefit of initial testing (e.g., Glover, 1989; Modigliani, 1976; Runquist, 1983, 1986). An early, parsimonious explanation for such results suggested that testing enhanced memory because engaging in a test provides reexposure to material that is successfully retrieved, thus increasing overall study time for tested information (cf. Carrier & Pashler, 1992; see also Thompson, Wenger, & Bartling, 1978, for a similar idea). However, this view has fallen out of favor, as testing effects are routinely observed in studies utilizing a restudy control condition, in which the retention of information subjected to testing is compared to that presented for an equal duration of restudy, thus equating overall exposure to materials between conditions (e.g., Carpenter & DeLosh, 2005, 2006; Carrier & Pashler, 1992; Kuo & Hirshman, 1996, 1997; Rowland & DeLosh, 2014b). Consequently, the testing effect seems tied to the act of testing itself.

Contemporary theoretical explanations of the testing effect are often specified in a general or abstract way or, alternatively, pertain explicitly to only a subset of testing phenomena. Thus, different theories are not, necessarily, mutually exclusive. One class of theories suggests that the testing effect arises from, and grows with, the difficulty or effort induced by initial retrieval. Although the terminology used may vary, this idea is captured by theories predicting that the magnitude of the testing effect should increase with the difficulty (e.g., Jacoby, 1978; Karpicke & Roediger, 2007), "completeness" (Glover, 1989, p. 392), depth (Bjork, 1975), or effort (e.g., Pyc & Rawson, 2009) of retrieval events, in all cases referring to the quality and intensity of processing that is induced by the retrieval attempt. Bjork and Bjork's (1992) new theory of disuse similarly captures the role of retrieval difficulty. The theory specifies that memories can be described in terms of their storage strength and retrieval strength. Storage strength refers to the degree to which a memory is durably established, or "well learned" (p. 42), whereas retrieval strength indexes the accessibility of a memory at a given point in time (i.e., how easily it can be retrieved). Critically, difficult tests (i.e., tests over information with relatively low retrieval strength) are thought to provide a more substantial benefit to storage strength when compared with easier tests. For the purposes of the present study, such theories are classified as retrieval effort theories, given the commonality of their most direct predictions. Regardless of the specific theory under consideration, there exists strong evidence in support of the role of retrieval difficulty in the testing effect.

Support for the influence of initial retrieval effort on the testing effect has come from studies showing that testing effects are larger following more difficult initial tests as indexed by diminishing cue support (e.g., Carpenter & DeLosh, 2006; Glover, 1989; Kang, McDermott, & Roediger, 2007) and increased delay between initial study and initial testing (e.g., Karpicke & Roediger, 2007; Modigliani, 1976; Pyc & Rawson, 2009; Whitten & Bjork, 1977). Pyc and Rawson (2009) showed that given multiple retrieval opportunities, a longer duration between subsequent retrievals led to better retention. Furthermore, the additive benefit of multiple retrieval opportunities was best fit by a negatively accelerating power function (i.e., each subsequent retrieval of an item was of lesser relative benefit), suggesting that as retrievals became easier, their mnemonic utility decreased (see also Bjork, 1975).

Although greater retrieval difficulty and effort can increase the magnitude of the testing effect, many theories generating such predictions do not specify the causal mechanisms at play. There are, however, a number of possibilities that are able to coexist with the basic assumptions specified by retrieval effort theories. Retrieval may increase the number of retrieval routes that are able to be effectively utilized at a later test (e.g., Bjork, 1975; McDaniel & Masson, 1985; see also Rowland & DeLosh, 2014a), promote distinctive or item-specific processing of tested information (e.g., Kuo & Hirshman, 1997; Peterson & Mulligan, 2013), or allow for elaboration of a target piece of information (e.g., Carpenter, 2009; see also Pyc & Rawson, 2010). For example, the *elaborative retrieval* hypothesis of the testing effect (Carpenter, 2009; Carpenter & DeLosh, 2006; see also Bjork, 1975) suggests that testing is beneficial as a direct result of the elaboration of a memory trace that results from engaging in retrieval operations. Carpenter (2009) proposed a mechanism for such elaboration, whereby engaging in a retrieval attempt produces an elaboration of the target of the memory search through the activation of semantic associates of the target. For example, having studied the cue–target pair ANIMAL–CAT, engaging in a recall test given the cue ANIMAL–? may induce a participant to generate plausible but erroneous candidates (e.g., dog, lion, kitten, etc.) prior to arriving at the target (cat). At a later memory test, the previously generated candidates may serve as retrieval cues for the target, thereby enhancing the likelihood of target retrieval. Because a similar elaborative structure is not likely to be generated for restudied items where a cue–target pair is represented in full form, a testing effect should obtain. Carpenter (2009) found support for the elaborative retrieval view by demonstrating that weakly associated cue–target pairs generate larger testing effects than strongly associated pairs, presumably because of the greater degree of elaboration potential for weakly associated pairs (i.e., a greater number of associates would be expected to become activated before arriving at the target during retrieval). Thus, more difficult initial testing conditions, which should encourage more extensive elaboration, should further enhance testing effects (e.g., given a difficult test, one would be expected to generate more erroneous candidates before reaching the target), in accordance with the predictions of retrieval effort theories as described above.

A similar, though more specific theory, the *mediator effectiveness* hypothesis (Pyc & Rawson, 2010), provides further specification as to the nature of elaboration that may occur during retrieval. The theory proposes that the testing effect can derive in part from more effective utilization of cue–target mediating information that arises as a result of testing, where "mediating information" refers to information of some sort (e.g., a word) that provides a link between a cue and target. Participants seem to spontaneously activate mediating information during testing in the form of semantic associates to cues (Carpenter, 2011). At final test, such mediating information is often remembered in response to the cue and can more effectively prompt recall of the target (Carpenter, 2011; Pyc & Rawson, 2010). Presumably the activation of semantic mediating information is more likely given difficult retrieval tasks when a more thorough search for the target is necessary, and thus the mediator effectiveness hypothesis specifies a plausible mechanism compatible with retrieval effort theories.

An additional point of interest, related to the importance of initial test difficulty, concerns the overall efficacy of recognition testing. If sufficiently difficult retrieval is necessary to induce testing effects, whether due to semantic elaboration or other means, an implication is that initial recognition testing may not always yield reliable or robust testing effects. In some cases, recognition testing has been found to benefit target retention (e.g., Mandler & Rabinowitz, 1981; Odegard & Koen, 2007; Roediger & Marsh, 2005; Runquist, 1983), although the effect does not reliably obtain (see, e.g., Glover, 1989; Runquist, 1983), especially when contrasted with a more conservative restudy control condition (e.g., Butler & Roediger, 2007; Carpenter & DeLosh, 2006; Kang et al., 2007; McDaniel, Anderson, Derbish, & Morrisette, 2007; Sensenig, 2010). If interpreted in the context of generate–recognize models of recall (e.g., J. R. Anderson & Bower, 1972), testing effects may result in some manner from the processing engaged in during the generation of candidate targets during a memory query (as could be expected by the elaborative retrieval hypothesis), rather than the subsequent recognition of candidates for output. Although there are potential indirect, "mediated" (Roediger & Karpicke, 2006a, p. 182) benefits of testing (e.g., tests can more effectively guide future learning following feedback; see Pyc & Rawson, 2011; see also Thompson et al., 1978), a retrieval-dependent direct benefit of testing should manifest as a greater magnitude testing effect following recall, compared with recognition initial testing (but see Little, Bjork, Bjork, & Angello, 2012, for one potential caveat).

Retrieval effort theories emphasize the conditions of the initial test. However, additional characterizations of the testing effect have drawn attention to the conditions of the final test. The bifurcation model (Kornell, Bjork, & Garcia, 2011; see also Halamish & Bjork, 2011), which is discussed more thoroughly in the Discussion, specifies that study (and restudy) provides a modest increase to memory strength, whereas testing provides a more substantial increase, but only to those items successfully retrieved. The framework makes no assumptions about the impact of initial test conditions in dictating the degree of strength increment, however. Instead, emphasis is placed on the difficulty of the final test. When the final test is more difficult, a testing effect will be more likely to occur, all else constant (and if so, it will be relatively larger in magnitude compared with an easier final test). The frequently observed finding that testing effects increase with longer retention intervals provides a base of support for the role of final test difficulty (e.g., Roediger & Karpicke, 2006a, 2006b; Toppino & Cohen, 2009; Wheeler, Ewers, & Buonanno, 2003).

Another contemporary theory of the testing effect explicitly calls attention to the importance of the *similarity* between initial and final testing conditions. The transfer appropriate processing (TAP) theory of the testing effect (drawing from C. D. Morris, Bransford, & Franks, 1977; see Bjork, 1988; Roediger & Karpicke, 2006a) suggests that the testing effect derives from the overlap in processing that occurs during initial and final testing. As a principle applied to many memory phenomena, TAP specifies that memory performance is positively related to the degree of overlap in processing that occurs during encoding and retrieval. Applied to the testing effect, TAP theory states that an initial test during learning, compared with restudy (or no test), induces a greater similarity with the type of processing that occurs at the final test. In one sense, initial tests grant practice at the task of retrieving information (hence the term *retrieval practice*).

TAP theory makes a clear prediction: The magnitude of the testing effect should depend on the degree of similarity between the initial test and the final test. The larger the degree of overlap in processing required for initial and final testing, the larger the benefit of the initial test. Empirically, TAP theory has received mixed support. Theoretically consistent results have been reported by Duchastel and Nungester (1982), who gave participants a passage to read, followed by either a short-answer test, a multiple-choice test, or no test. Following a 2-week delay, a final test was given with both multiple-choice and short-answer questions. Both test groups outperformed the no-test group. However, of key interest, performance on final multiple-choice questions was highest for the group that received initial multiple-choice testing. McDaniel and Fisher (1991, Experiment 2) found that successful initial testing on factual questions led to better performance on a later test using identical questions compared with rephrased versions of questions, suggesting that the greater overlap between learning and assessment benefited performance. Similarly, Johnson and Mayer (2009) had participants view a multimedia presentation about lightning, followed by either a restudy opportunity, a practice retention test (asking the participant to describe "how lightning works"), or a practice transfer test (asking questions whose answers could be derived from the content of the presentation). On a delayed test 1 week later, participants were given either a retention test (asking "how lightning works") or a transfer test, similar to the practice-transfer test but with additional questions. The results showed a crossover interaction between practice test type and final test type, such that performance was highest when initial and final test types matched, as predicted by TAP theory.

Despite such evidence that is compatible with TAP theory, a body of research has also produced results problematic for the theory. Carpenter and DeLosh (2006) fully crossed both initial and final test types (free recall, cued recall, and recognition) and found that regardless of the format of the final test, free recall initial testing yielded the best performance (see also Glover, 1989). Similarly, Kang et al. (2007, Experiment 2) found that initial short-answer testing (with feedback) led to greater final multiple-choice test performance than did initial multiple-choice testing, a finding that contrasts with Duchastel and Nungester (1982) and is at odds with TAP theory. Other studies show that more closely matching the characteristics of the initial and final tests can yield no benefit or even suboptimal relative performance (e.g., Carpenter & DeLosh, 2006; Rohrer, Taylor, & Sholar, 2010). Thus, TAP theory, considered in isolation, has unclear explanatory power for the testing effect.

Nonetheless, TAP theory is commonly presented as a primary, contemporary, theoretical contender for explaining the testing effect (see, e.g., Bouwmeester & Verkoeijen, 2011; Halamish & Bjork, 2011; Johnson & Mayer, 2009; Karpicke & Roediger, 2007; Roediger & Karpicke, 2006a, 2006b; Sumowski, Chiaravalloti, & DeLuca, 2010), and thus remains influential in guiding and framing empirical investigations and patterns of results. Perhaps more important, TAP theory is not inherently at odds with retrieval effort theories or the bifurcation model, in that it is possible that initial retrieval conditions and/or final retrieval conditions, in addition to the relative match in processing between learning and assessment, jointly contribute to the testing effect.

Meta-analysis provides a means by which to adjudicate between the existing classes of theories that have been applied to the testing effect. Retrieval effort theories typically emphasize the role of initial test conditions; the bifurcation model draws attention to the conditions of the final test; and TAP theory notes the importance of the interaction between initial and final test conditions. As such, three moderator variables are included in the meta-analysis: initial test type (recognition, cued recall, and free recall), final test type (recognition, cued recall, and free recall), and initial–final test match (same, different).

Boundary Conditions

Although there is relatively little theoretical work on the testing effect, there are a variety of empirical issues that have implications for theories of the testing effect, yet remain unresolved. Here I outline three groups of factors that have been proposed as boundary conditions of the testing effect and indicate the relevant study characteristics evaluated in the present meta-analysis that may help inform the factors. The factors to be discussed are the impact of experimental design, the moderating influence of retention interval, and the impact of stimuli and cue characteristics.

Experimental Design Issues

A large number of memory effects that have been characterized as reliable are sensitive to certain manipulations of experimental design (e.g., the generation effect, the word frequency effect, and the bizarreness effect, to name a few; see McDaniel & Bugg, 2008; see also Erlebacher, 1977). Two factors have gathered much attention: the use of within- compared to between-participants manipulations of the variable of interest and the use of pure compared to mixed-list designs (i.e., whether lists of stimuli are composed of items intermixed from each condition or not). Studies that have identified moderating effects of design and list composition on various phenomena have spurred theoretical development, in some cases leading to cohesive frameworks from which to describe seemingly disparate memory effects (e.g., McDaniel & Bugg, 2008). As such, it is useful to examine these experimental design factors in the context of the testing effect, and in doing so, compare and contrast the testing effect with other, related memory effects. As an illustration, consider the generation effect.

The testing effect is often treated as synonymous with a related, retrieval-based memory phenomenon: the generation effect. The latter refers to the finding that when information is learned by generation, retention is enhanced relative to learning through reading without generation. For example, generating a word from a fragment (ANIMAL–C_T) often leads to better retention than simply reading a word (ANIMAL–CAT). A key contrast with the testing effect is that generation requires retrieval from semantic memory, whereas testing requires retrieval from episodic memory (e.g., retrieving a target from a specific, previously learned list of words). Consequently, differences found between testing and generation may help distinguish between the role of semantic and episodic information as they contribute to the testing effect.

Along these lines, Karpicke and Zaromb (2010) demonstrated an empirical difference between the generation effect and testing effect in a study that manipulated "retrieval mode"—whether participants were explicitly instructed to retrieve targets from a previously learned list of material or to instead generate them from semantic memory. Explicit instructions to engage in intentional retrieval (i.e., testing) yielded a greater benefit to memory retention than initial generation, suggesting that at least the magnitude of the effects may differ, all else constant. It remains unclear, however, whether factors that have been found to moderate the generation effect but have not been applied to the testing effect, operate comparably. Generation effects tend to be larger in within-participant designs (in which mixed lists are commonly used) than between-participants designs (in which pure lists are commonly used; Bertsch, Pesta, Wiscott, & McDaniel, 2007). In a list learning paradigm, a *pure* list refers to a design in which each list within a study is composed entirely of items of a single condition (e.g., all generate condition items, or all read condition items). In contrast, a *mixed* list refers to a design in which every list is composed of items from both conditions (e.g., half generated and half read items, intermixed). The generation effect appears robust when mixed lists are used but is mitigated, null, and in some cases reverses to a generation disadvantage (relative to a read condition), when pure lists are employed (e.g., Nairne, Riegler, & Serra, 1991; Serra & Nairne, 1993; Slamecka & Katsaiti, 1987).

Of theoretical relevance, one framework used to explain the list composition discrepancy in the generation effect suggests that generating information may enhance item-specific processing at the expense of relational processing of serial order information, with the converse true for reading (see Nairne et al., 1991; Serra & Nairne, 1993). Given a pure list, read condition items attain a relative enhancement in the encoding of serial order that can later be used to guide and organize recall, presumably overcoming the item-specific processing advantage for generated items. Given mixed lists, however, the presence of generated items intermixed with read items disrupts order encoding for all items, unmasking the individual-item advantage for generated items. This account has been generalized and extrapolated to a variety of memory phenomena (see McDaniel & Bugg, 2008), and thus might be expected to apply to the testing effect.

In the testing effect literature, both within-participant and between-participants designs, in addition to pure and mixed lists, have been used to examine testing effects (e.g., Carrier & Pashler, 1992; Karpicke & Zaromb, 2010; Rowland et al., 2014). Studies suggest that the testing effect is reliable across all of these experimental designs. It is not clear, however, whether design has a reliable impact on the magnitude of the testing effect. In addition, although Karpicke and Zaromb (2010) found that testing, like generation, impaired order memory, they also found that testing effects still obtained under conditions where generation effects did not. Other recent investigations have found little evidence for list-type interactions with the testing effect (Rowland et al., 2014), and positive effects of testing on relational processing and organization in memory over time (e.g., Congleton & Rajaram, 2011, 2012), drawing from measures of both semantic (i.e., categorical) clustering of items and subjective (i.e., not inherently categorical) grouping of items (Zaromb & Roediger, 2010). As such, the fact that the testing effect is an episodic retrieval task may provide a means by which episodically linked information in memory can provide an organizational structure that benefits later recall, even if order encoding, specifically, is disrupted. In sum, this meta-analysis provides an opportune chance to investigate whether the testing effect consistently responds to changes in design and list composition in a manner similar to the generation effect and may thus provide an avenue for further development on the contribution of both semantic and episodic-bound information to the testing effect. As such, experimental design and list blocking were included as moderator variables in the meta-analysis.

Retention Interval as Moderator

Although the testing effect appears to be a robust phenomenon, the duration of the retention interval has been shown, in many cases, to moderate the effect. The retention interval refers to the duration of time between the end of acquisition (i.e., initial study and restudy or initial testing) and the start of the final test. There is substantial agreement in the literature that testing effects emerge following long retention intervals (of days or weeks). However, many researchers have observed that testing effects are of lesser magnitude, null, or even reverse to restudy advantages, at short retention intervals on the order of minutes (e.g., Bouwmeester & Verkoeijen, 2011; Congleton & Rajaram, 2012; Roediger & Karpicke 2006b; Toppino & Cohen, 2009; Wheeler et al., 2003). In contrast, numerous studies have demonstrated reliable testing effects at short retention intervals similar or identical to those used in the studies cited above (see, e.g., Carpenter & DeLosh, 2005, 2006; Karpicke & Zaromb, 2010; Kuo & Hirshman, 1996; Rowland & DeLosh, 2014b). Thus, the form of the interaction between testing and retention interval (i.e., whether it is ordinal or disordinal) has not been firmly established, primarily because of highly variable findings following short retention intervals.

Of theoretical relevance, numerous characterizations of the testing effect have been motivated in large part by attempting to explain the testing by retention interval interaction (e.g., Congleton & Rajaram, 2012; Verkoeijen, Bouwmeester, & Camp, 2012). These accounts typically suggest that testing and study induce qualitatively different types of processing, thereby providing a means of explaining the disordinal interaction of retention interval and learning method that has been observed in some studies. If this is the case, future theoretical development of the testing effect would benefit from further elucidation of the precise differences in processing resulting from study versus testing. Alternatively, the interaction may result, at least in part, from characteristics of the experimental design employed (see, e.g., Kornell et al., 2011; Rowland & DeLosh, 2014b) rather than (or in addition to) qualitative differences in processing between testing and studying. Moreover, of theoretical note, TAP theory has difficulty explaining why study should yield superior retention compared with initial testing following short intervals. That is, regardless of the retention interval, testing should always induce a greater similarity in processing between learning and assessment when compared to restudy. It is therefore of both empirical and theoretical importance to assess the reliability of the testing effect as a function of retention interval in order to evaluate existing theoretical accounts, as well as guide future theoretical development of the testing effect. Retention interval was thus treated as a moderator variable in the meta-analysis.

Variability Across Materials

Testing has been demonstrated to benefit retention in studies using highly variable materials and testing conditions. This is one reason why it has been embraced by cognitive scientists as a recommendation for practical application (see, e.g., Roediger & Karpicke, 2006a), with laboratory research on the testing effect increasingly becoming disseminated to audiences outside the field (e.g., Karpicke & Grimaldi, 2012; Rawson & Dunlosky, 2012; see also Agarwal, 2012, for a relevant brief commentary on the contemporary effort to bridge cognitive research and educational practice). In light of studies that have implemented testing procedures during learning in simulated (Butler & Roediger, 2007; Campbell & Mayer, 2009) and actual classroom contexts (e.g., Carpenter, Pashler, & Cepeda, 2009; Gingerich et al., in press; McDaniel, Agarwal, Huelser, McDermott, & Roediger, 2011; McDaniel et al., 2007; Roediger, Agarwal, McDaniel, & McDermott, 2011; see also Bangert-Drowns et al., 1991), it appears that the general findings reported from laboratory studies generalize to the classroom (i.e., testing appears beneficial). It is not clear, however, if the format or characteristics of learned materials have an impact on the absolute magnitude of the testing effect. Furthermore, there are a number of theoretically motivated characterizations of the testing effect that, in some cases, make either explicit or plausible predictions that the testing effect should be sensitive to characteristics of stimuli being learned and the cues provided during retrieval attempts.

Verbal materials are the most commonly used types of stimuli in testing effect studies and are typically presented as lists of single words (e.g., Carpenter & DeLosh, 2006), paired associates (e.g., Pyc & Rawson, 2010), or prose passages (e.g., Roediger & Karpicke, 2006b). Studies using each of these classes of materials have consistently yielded reliable testing effects, although no study has investigated the possibility of their relative impact on the magnitude of the effect. Of particular relevance to this meta-analysis, recent theoretical proposals have made the argument that testing, compared with study, may differentially influence the processing of materials presented during learning.

One idea, drawing from fuzzy trace theory (Reyna & Brainerd, 1995), suggests that studying encourages "verbatim" processing (i.e., the processing of surface characteristics of stimuli), whereas testing encourages "gist" processing (i.e., the processing of abstracted, semantic characteristics or common features between stimuli) of previously studied information (Bouwmeester & Verkoeijen, 2011; Verkoeijen et al., 2012). This view is consistent with findings showing that tests can increase the occurrence of false memories for semantically related lures when given lists with a common semantic theme (e.g., McDermott, 2006). Drawing from this characterization, the presence or absence of an underlying semantic theme within a set of stimuli may influence the magnitude or presence of the testing effect. In the case of semantically unrelated sets of stimuli, testing (relative to study) may enhance processing of gist or other abstracted conceptual information (Verkoeijen et al., 2012), which can, at a later assessment, function as effective cuing information. However, given sets of stimuli in which a semantic gist or theme is inherently present (e.g., Deese–Roediger–McDermott [DRM] lists; see Roediger & McDermott, 1995), the beneficial effect of enhanced gist processing from testing may be somewhat redundant. That is, gist information may be readily apparent, and thus could serve as an effective retrieval cue at a final test even without prior testing. Thus, restudying may be of equal or greater effectiveness at promoting retention through its relative enhancement of verbatim processing (see Delaney, Verkoeijen, & Spirgel, 2010). Note that alternative predictions may arise depending on study characteristics (such as the retention interval or the cue support available at final test; for more thorough and nuanced discussion of such possibilities, see Bouwmeester & Verkoeijen, 2011; Verkoeijen et al., 2012). However, assessing the impact of the relationship between to-be-learned materials on the testing effect seems an ideal area for exploratory analysis, to which meta-analysis is well suited.

Given somewhat similar core assumptions to the fuzzy trace view described above, an alternative outcome can be predicted. Testing appears to encourage organizational processing in memory (Congleton & Rajaram, 2011, 2012; Zaromb & Roediger, 2010), such that conceptually related information within a given stimulus set is more cohesively grouped during output, with the grouping enduring over time relative to learning with study only. This may reflect a relative enhancement in relational over item-specific processing following testing (Congleton & Rajaram, 2011, 2012; but cf. Karpicke & Zaromb, 2010; Peterson & Mulligan, 2013; see also Hunt & McDaniel, 1993, for elaboration on item and relational processing). Similarly, testing, more so than restudy, can facilitate the retention of semantically related but untested information from an initial learning period under certain circumstances (e.g., Chan, 2009, 2010; Chan, McDermott, & Roediger, 2006; Cranney, Ahn, McKinnon, Morris, & Watts, 2009; Rowland & DeLosh, 2014a). If testing promotes retention for both the target information retrieved and other semantically related material encoded during initial learning, the presence of an underlying conceptual or semantic structure to stimuli could be exploited, thereby strengthening the testing effect. However, the facilitation effect does appear to be sensitive to the level of integration and coherence embedded in the stimuli (Chan, 2009; see also Little, Storm, & Bjork, 2011; but cf. Rowland & DeLosh, 2014a). In the absence of such integration, testing can harm retention of semantically related information (i.e., retrieval-induced forgetting can occur; see M. C. Anderson, 2003; Storm & Levy, 2012, for reviews). Given such characterizations of testing, the above-mentioned classes of verbal materials can then, in somewhat different terms, be classified as containing either semantically unrelated stimuli (e.g., unrelated word lists), unstructured but semantically themed stimuli (e.g., categorized lists), or structured, semantically rich stimuli (e.g., prose passages). Thus, such classification of stimulus interrelation was treated as a moderator in the meta-analysis.

A second characteristic of interest concerning the stimuli employed in studies of the testing effect is the relationship, if any, afforded between initial test cues and targets. For example, one instantiation of the elaborative retrieval hypothesis suggests that taking a test encourages the production of semantic mediators, that is, concepts that create a link between cue and target (Carpenter, 2011; Pyc & Rawson, 2010). Carpenter (2009) demonstrated that the degree of semantic relatedness between cue and target influenced the magnitude of the testing effect, perhaps because weaker relationships induce a greater degree of elaboration, and potentially more semantic mediators that can effectively guide retrieval to a target. A related prediction, however, is that cue–target relationships that allow for semantic elaboration (i.e., materials in which both a cue and a target inherently contain semantic information) should benefit to a larger degree by testing than materials lacking the potential to utilize semantic mediation (for variations on this idea, see Kang, 2010; Sensenig, Littrell-Baez, & DeLosh, 2011). For example, given a cue–target pair of a face and name, neither the cue nor the target clearly contains inherent semantic information, and thus would not as readily benefit from test-induced semantic elaboration. Even so, testing effects have been reported in the literature with materials carrying limited or effectively no inherent semantic content, such as names cuing names (e.g., P. E. Morris, Fritz, Jackson, Nichol, & Roberts, 2005, Experiment 1), faces cuing names (e.g., Carpenter & DeLosh, 2005; P. E. Morris et al., 2005, Experiment 2), words cuing unfamiliar symbols (e.g., Kang, 2010), unfamiliar symbols cuing words (e.g., Coppens, Verkoeijen, & Rikers, 2011), fragments of names cuing full names (e.g., Sensenig et al., 2011), and unknown foreign language words cuing known words (e.g., Carpenter, Pashler, Wixted, & Vul, 2008; Carrier & Pashler, 1992; Toppino & Cohen, 2009). Given that the testing effect still emerges under such circumstances, a theoretically motivated question of interest is whether the potential for semantic elaboration between cue and target can increase the magnitude of testing advantages.

In sum, the specific characteristics of the targets and cues provided during learning may be of theoretical and practical importance to the testing effect. Despite the apparent robustness of the testing effect across stimuli and cuing procedures, meta-analysis provides an ideal opportunity to assess what stimuli and cuing features, if any, have an impact on the magnitude of the testing effect. Concerning the role of stimuli characteristics, three additional moderators are considered in the meta-analysis: the format of stimuli used (e.g., single words, paired associates, prose), the presence of conceptual or semantic relatedness between target stimuli, and the potential for semantic elaboration between cues and targets during initial testing.

The Present Meta-Analysis

If described by analogy to an experiment, a meta-analysis treats individual studies (or samples of participants within studies) as participants, effect sizes (based on Cohen's *d* for the present analysis) as data points, and methodological characteristics as moderators. The effect sizes in the present meta-analysis were derived by comparing the retention of test condition information to the retention of restudy condition information. Positive effect sizes indicate a test advantage, whereas negative effect sizes indicate a restudy advantage on the final test. The selection of studies for the present meta-analysis, the coding of moderator variables, the calculation of effect sizes, and a description of the analyses employed are described below.

Method

Search Strategy and Coding Procedures

All literature searches and coding procedures were carried out by the author except where described otherwise. Studies to be considered for inclusion in the meta-analysis were gathered by means of three primary methods. First, electronic searches of scientific publication databases were conducted (PsycINFO, Google Scholar, Dissertation Abstracts) using combinations of the following terms: *test**, *effect*, *retrieval*, *recall*, *practice*, and *memory*. After locating indexed studies, forward and reverse citation searches were performed for existing review articles (e.g., Rawson & Dunlosky, 2011; Roediger & Karpicke, 2006a) to identify additional studies not captured by the database searches. Finally, requests for unpublished data were made to a number of researchers with recent publications related to the testing effect. In some cases, researchers who were directly contacted provided referrals to other researchers affiliated with their labs, thus broadening the search for unpublished studies. All studies included in the meta-analysis were gathered by March 2013. No date range constraint was applied in the literature search, though the earliest studies assessed for inclusion were published approximately one century ago (e.g., Abbott, 1909). Note, however, that research on retrieval effects in memory using paradigms resembling those examined in the meta-analysis largely emerged in recent decades. Even so, studies published from all available dates were assessed following the same method and criteria. The initial search yielded 308 published and 23 unpublished studies that were deemed potentially relevant to the topic based on examinations of titles and abstracts.

The full text of those studies deemed potentially relevant to the topic under investigation were then examined to determine their relevance to the topic and their fit to the general methodological framework as specified in the introduction (see Scope of the Present Meta-Analysis). In order to be considered, studies needed an initial learning phase in which all information was studied for a similar duration, a test versus restudy manipulation taking place during an intervening phase, and a final assessment over the learned material ($k = 240$ studies removed). Studies that met these criteria and were judged as potentially suitable to the topic were assessed against the following criteria for inclusion. Studies must have (a) assessed memory performance for the same information that was initially tested and/or restudied ($k = 4$ studies removed); (b) treated all information equally in both test and restudy conditions ($k = 3$ studies removed); (c) not instructed participants to restrict output of certain items at the final assessment ($k = 1$ study removed); (d) not assessed participant learning of information for a real or simulated class in which they were enrolled ($k = 13$ studies eliminated; but see Bangert-Drowns et al., 1991, for a relevant meta-analysis); and (e) not utilized a clinical participant population ($k = 1$ study removed). In addition, studies from which there were insufficient data for calculating effect sizes (i.e., reported neither in text nor graphically in figures) were excluded from the meta-analysis ($k = 8$; note that in all cases these studies were published at least 15 years prior, and thus missing data were not requested).

Effect sizes were derived from independent samples of participants within studies such that no two effect sizes included in the analysis utilized data contributed from overlapping subsets of participants. The outcome measure derived from each study was memory performance, defined as the proportion of previously learned items correctly remembered on the final assessment as a function of learning procedure (i.e., testing vs. restudy). For studies in which the same sample yielded multiple measures of memory performance (e.g., multiple, identical final tests were administered on unique subsets of initially studied information; e.g., Carpenter et al., 2008, examined memory for unique subsets of studied information at varied retention intervals for each participant), one performance measure was randomly selected to be included in the analysis. As such, in all cases, effect sizes were calculated by examining recall or recognition performance of initially tested versus restudied information. One hundred and fifty-nine effect sizes drawing from data reported in 61 studies, published or otherwise reported from 1975 through 2013, met inclusion criteria and were used in the analysis. Studies are reported in Appendix A.

Coding protocols were determined a priori (e.g., identification of moderators and levels of interest) based on theoretical interest, with the exception, in a few cases, of select categories within moderators either being collapsed together or expanded into multiple categories based on examining the literature after the onset of the literature search. Such cases are noted below when describing moderator variables. Coding of information derived from studies was completed by the present author for all studies included in the meta-analysis. However, a random 20% of studies were provided to an independent coder, trained by the author, with experience in conducting testing effect research following the general paradigm used for those studies under investigation. Interrater reliability in coding study categorical variables was high (in each case $k \geq 0.92$), and discrepancies were resolved through discussion.

Moderator Variables

In addition to determining an estimated mean effect size, meta-analysis can be used to estimate the impact of study characteristics on effect sizes. Studies meeting inclusion criteria were coded with respect to the following study characteristics and design components to be used in moderator analyses.

Publication status. A concern to all meta-analyses is the possibility of publication bias, in which the published literature may misrepresent the true magnitude of the effect of interest. Publication status was thus coded as a categorical variable with two levels: published and unpublished. Theses and dissertations were coded as unpublished. The issue of publication bias is explored further in the Discussion.

Sample source. The literature on testing effects has almost exclusively used samples of undergraduate college students. However, a few recent studies have used samples deviating from this norm, utilizing Internet solicitation (e.g., Carpenter et al., 2008), and older adult (e.g., Bishara & Jacoby, 2008) to high-school (e.g., P. E. Morris et al., 2005) and younger (e.g., Bouwmeester & Verkoeijen, 2011; Rohrer et al., 2010) age samples. Because of the very small number of effect sizes attained from any single sample source group other than college students, sample source was coded as a categorical variable with two levels: college and other, the latter of which resulted from collapsing more diversely coded sample sources after the onset of coding.

Design. Study design was coded as a categorical variable with two levels: between participants and within participant, referring to the test versus restudy manipulation.

Stimulus type. To assess the impact of the type of to-be-learned materials on the testing effect, stimulus type was coded as a categorical variable with four levels: single words, paired associates, prose, and other. The "other" category represented five effect sizes drawn from studies using maps (Carpenter & Pashler, 2007; Rohrer et al., 2010), multimedia presentations (Johnson & Mayer, 2009), and obscure facts (Carpenter et al., 2008). These were represented in a single category due to the low number of effect sizes available in any given subcategory, a modification to the coding procedure made after the onset of coding.

Stimulus interrelation. The testing effect is often studied with either structured prose passages or unrelated verbal materials. However, a few investigations have utilized stimuli that are not captured in these two classes of materials, such as DRM lists (e.g., Verkoeijen et al., 2012) or other unstructured lists representing categorical or semantically themed items (e.g., Congleton & Rajaram, 2012). The relationships between target information to be learned was coded as a categorical variable with four levels: prose, categorical (i.e., unstructured, semantically themed materials), unrelated (i.e., semantically unrelated materials), and other, with the "other" group capturing materials not clearly belonging to the three primary groups (e.g., maps).

List blocking. For those studies utilizing a list learning procedure, a moderator of interest is the structure of the lists employed. Blocked list designs are those in which all test condition and restudy condition items are presented in segregated lists (or between participants). Alternatively, mixed lists are those in which both test and restudy condition items are intermixed within lists. As such, list blocking was coded as a categorical variable with three levels: mixed, blocked, and other. Note that some studies (e.g., Brewer & Unsworth, 2012) used a single list representing both tested and restudied information in a blocked fashion, such that, for example, all test items appeared first, followed by all restudy items. In such cases, list blocking was coded as blocked, reflecting the blocking of items by condition. The level of other was included to represent studies in which the materials were not organized in a list-like fashion (e.g., prose passages).

Initial test type. The format of the initial test was coded as a categorical variable with three levels: free recall, cued recall, and recognition.[1] Although some types of recognition testing procedures may encourage recall (see, e.g., Little et al., 2012), all studies using multiple-choice tests were coded as recognition tests.

Number of initial tests. The number of initial tests that a test condition item received during learning was coded as a continuous variable (range: 1–5).

Initial test lag. The delay, in seconds, between initial study and the initial test was coded as a continuous variable. Initial test lag was determined by calculating the average time between original exposure and first test across test condition items and, as such, included any lag imposed by additional items in a given list. For studies that did not report explicit timing information, or that allowed self-paced study or testing, an estimate was made based on the available information reported and by the timing used in similar procedures described in the literature (e.g., other experiments reported in the same study). If no estimate could be made with high confidence (e.g., due to high timing variability between participants or trials), the initial test lag moderator was not coded and the corresponding study was dropped from the moderator analysis (this was the case for three studies: Fadler, Bugg, & McDaniel, 2012; Hinze & Wiley, 2011; P. E. Morris et al., 2005, Experiment 1). In the case of studies where participants were permitted to study the entire body of material concurrently in the test condition (e.g., prose passages), average initial test lag was estimated as half of the time provided for initial study, plus any experimenter imposed lag between initial study and initial test.

Retention interval. The duration between the end of the acquisition period (i.e., initial study and restudy or test) and the beginning of the final memory assessment, in minutes, was coded in two ways. As a continuous moderator, the duration of the retention interval, in minutes, was subjected to a logarithmic transformation. An additional categorical moderator analysis was conducted, in which studies were separated into two groups: those with retention intervals less than 1 day and those with retention intervals 1 day or longer.

Initial test cue–target relation. The potential for and nature of a semantically elaborative relationship between the retrieval cue provided at initial testing and the target to be retrieved was coded as a categorical variable with five levels: nonsemantic, semantic unrelated, semantic related, same (i.e., recognition testing where the target is provided as a cue), and none (i.e., free recall testing). After identifying studies in the same and none categories, the remaining studies were first partitioned according to the potential for semantic elaboration between cues and targets. Studies were only defined as having potential for semantic cue–target elaboration if both the cue and the target carried inherent, known semantic meaning (e.g., word–word pairings, but not names, face–name pairs, symbol–word pairs, or foreign language translations). As such, even though some procedures may have allowed for the use of relational semantic processing for

[1] An additional test type category, matching, was to be included to reflect studies using a matching task (Rohrer, Taylor, & Sholar, 2010; Wartenweiler, 2011). However, only three effect sizes were derived from such studies, and thus matching tests were coded as recognition tests.

some participants given items without inherent semantic content (e.g., name learning), these studies were defined for present purposes as *nonsemantic*. Those studies that included both cues and targets carrying inherent semantic content were then subdivided according to the nature of the existing semantic relationship between cues and targets. Studies indicating that cue–target pairs were semantic associates to any degree (e.g., weakly, strongly, or related but of unspecified strength; e.g., CAT–LION) were coded as *semantic related*, as a more fine-grained analysis as a function of the degree of semantic relatedness (e.g., unrelated vs. weakly related vs. strongly related) yielded undesirably small category sizes. Studies in which both cues and targets had inherent semantic content but were themselves not semantically related (e.g., CAT–BROCCOLI) were coded as *semantic unrelated*. Thus, the initial test cue–target relationship moderator classified studies according to their potential for semantic elaboration between cues and targets, and if present, the existing nature of the cue–target semantic relationship.

Final test type. The format of the final memory assessment was coded as a categorical variable with three levels: free recall, cued recall, and recognition.

Initial–final test match. The match between initial and final test formats (given the possibilities of cued recall, free recall, and recognition) was coded as a categorical variable with two levels: same and different.

Feedback. Providing feedback can enhance final test performance for test condition items (e.g., Kang et al., 2007). As such, the presence of feedback after an initial test (as well as similar additional restudy) was coded as a categorical variable with two levels: yes and no. Note that a study was coded as yes for feedback if the tested material was reexposed after any initial test (e.g., in the form of alternating study–test cycles; see, e.g., Karpicke & Blunt, 2011; Zaromb & Roediger, 2010, Experiment 1).

Retrievability and reexposure. Many of the positive effects of testing may be limited to circumstances in which the retrieval attempts are successful (see, e.g., Jang, Wixted, Pecher, Zeelenberg, & Huber, 2012; Rowland & DeLosh, 2014b). Furthermore, in the absence of feedback, initial test performance serves as a proxy for the amount of reexposure to test condition items (given that unsuccessfully retrieved items are not reexposed during testing). As such, to help elucidate the effect of retrieval success and reexposure to test condition items, initial test performance was considered in tandem with feedback. Retrievability and reexposure were coded as a categorical variable. Studies were grouped first by whether they included feedback (as defined in the same manner as the feedback moderator, discussed above). Studies that did not provide feedback were grouped into categories according to their observed initial test performance. In all, the five levels were coded as feedback, no feedback > 75, no feedback 51–75, no feedback ≤ 50, and unknown, where the numbers following a given no-feedback level indicate the range of initial test performance (out of 100%). Note that studies that included feedback and reported initial test performance were still categorized into the feedback group, regardless of initial test performance. Initial test performance was coded as it was reported. If multiple initial test scores were reported, the last value was used.

Effect Size Calculations

In a meta-analysis, each data point is represented as an effect size. In the present study, each effect size indicates the standardized difference between test condition material and restudy condition material in the form of proportion correct on the final memory assessment. This standardized difference (Cohen's *d*) was calculated for studies with independent samples (i.e., between-participants designs), when means and standard deviations were reported, as the difference between means divided by the pooled standard deviation of the test and restudy conditions:

$$d = \frac{M_T - M_R}{\sqrt{\frac{(n_T - 1)(s_T)^2 + (n_R - 1)(s_R)^2}{n_T + n_R - 2}}}, \quad (1)$$

where subscript T values indicate the test condition, subscript R values indicate the restudy condition, M is mean proportion correct for a given condition, n is the sample size for a condition, and s is the standard deviation for a condition.

The pooled within-group standard deviation component of Equation 1 (i.e., the denominator) is recommended to be replaced with the standard deviation of difference scores when computing d using matched data (i.e., data from within-participant designs; Borenstein, 2009). However, such a calculation requires the correlation between test and restudy condition scores to be known, a value that is seldom, if ever, reported in the testing effect literature. Thus, a correlation of .5 was imputed for studies using within-participant designs, yielding an algebraically equivalent effect size calculation as in Equation 1.

For studies with independent samples not reporting sufficient information for use of Equation 1, t values were used to calculate d according to

$$d = t\sqrt{\frac{n_T + n_R}{n_T n_R}}. \quad (2)$$

When only p or appropriate F values were reported, they were first converted to equivalent t values before being used in Equation 2. Dunlap, Cortina, Vaslow, and Burke (1996) provide an alternative calculation for d for matched designs to prevent overestimates of d by the use of Equation 2. For such circumstances, the following was used to calculate d:

$$d = t\sqrt{\frac{2(1 - r)}{n}}, \quad (3)$$

where n is the sample size and r is the correlation between test and restudy scores (set to .5).

Cohen's d produces a slight overestimate of true effect size given small samples. To correct for this bias, the correction factor, J (a high-accuracy correction approximation from Hedges, 1981),

$$J = 1 - \frac{3}{4(df) - 1}, \quad (4)$$

is applied to d, where df is $n_T + n_R - 2$ for between-participants designs and $n - 1$ for within-participant designs. This yields the adjusted effect size, Hedges's g:

$$g = J * d. \quad (5)$$

All analyses reported were conducted using effect sizes as measured by Hedges's g, which can be interpreted in the same way as d, with positive g values indicating positive testing effects (i.e., higher final test performance for test condition information compared to restudy condition information).

Method of Analysis

Analyses were carried out using Comprehensive Meta-Analysis 2.0. Effect sizes in the form of Hedges's g were statistically combined using a random-effects model (see Hedges & Vevea, 1998).[2] Although fixed-effect models are commonly used in psychological meta-analyses (Schmidt, Oh, & Hayes, 2009), there are compelling theoretical and statistical reasons to prefer the use of a random-effects model (see, e.g., Hunter & Schmidt, 2000; Schmidt et al., 2009). A fixed-effect model assumes that all studies included in the analysis provide estimates of a single, constant (i.e., fixed) population effect size, with any observed variability due to sampling error alone. Alternatively, a random-effects model assumes that the analyzed effect sizes are drawn from a distribution of true effect sizes, with the true effect distribution representing a universe of both existing and nonexisting but potentially conducted studies. Thus, observed variance in analyzed effect sizes represents both sampling error and between-study variation in true effect sizes when using a random-effects model. Only a random-effects model allows the results of the meta-analysis to be validly generalized to the larger "population" of potential studies (e.g., to address the general question "What is the effect of testing versus restudy on retention?"), whereas the fixed-effect approach only allows one to infer about the specific sample of studies included in the analysis (see Hedges & Vevea, 1998), and thus offers little statistically justified practical utility in informing some of the questions asked in the present study.

Each effect size was weighted in relation to its sample size (i.e., effect sizes derived from large samples were given more weight in the analyses than small sample effect sizes). The weight, w, was equal to the inverse of the unconditional variance, v_U, of g (i.e., $w = 1/v_U$):

$$v_U = v_c + (tau)^2, \quad (6)$$

where tau^2 represents the between-study variability (i.e., heterogeneity), calculated using the method of moments, and v_c is the conditional variance of g (i.e., the sampling error variance):

$$v_c = J^2\left(\frac{1}{n_T} + \frac{1}{n_R} + \frac{d^2}{2(n_T + n_R)}\right) \quad (7)$$

for independent groups and

$$v_c = J^2\left(\left(\frac{1}{n} + \frac{d^2}{2(n)}\right)2(1 - r)\right) \quad (8)$$

for matched groups, with r imputed as .5, thereby nullifying the rightmost term.

A number of different sets of information are reported in the Results section. First, the primary analysis is reported. The presence of heterogeneity in effect sizes beyond that accountable to sampling error alone was tested by use of the homogeneity statistic, Q. Tau^2 indicates the between-study variance component that is tested against 0 through the Q statistic. A significant Q test indicates the presence of between-study heterogeneity in effect sizes beyond that expected by sampling error alone. The magnitude of the homogeneity test can be reported with the I^2 statistic (Higgins, Thompson, Deeks, & Altman, 2003), derived from Q, which indicates the percentage of total variance observed in the sample of effect sizes due to between-study heterogeneity rather than sampling error. Categorical moderator analyses, conducted as mixed-effect analyses (i.e., partitioning studies into levels of a given moderator and combining studies within levels using a random-effects model), are then described, which report both point estimates for g and 95% confidence intervals (CIs), in addition to tests of homogeneity between moderator levels, represented by the Q_B statistic (a common tau^2 was not assumed across groups in such analyses). The Q_B statistic is interpreted similarly to the F statistic reported for a one-way analysis of variance, in that a significant Q_B indicates that not all levels of the moderator being tested are of reliably equal effect size. Last, continuous moderator analyses are presented. Continuous moderators were each analyzed by fitting a meta-regression model using the method of moments, where a slope that reliably differs from 0 indicates a relationship between a moderator and the magnitude of the testing effect.

Analyses were carried out on a primary, full data set ($k = 159$), along with a supplementary high-exposure ($k = 92$) data set. The full data set included all studies selected for the meta-analysis, whereas the high-exposure data set included only those studies that either provided feedback or yielded greater than 75% initial test performance. The high-exposure data served two primary functions. First, this data set reduces the conservative bias that is present in the full data set, given that all studies included in the present analysis provided full reexposure to control condition items (through restudy), whereas only those test items successfully recalled (or followed by feedback) were reexposed. In addition, the high-exposure data set was used to provide additional evidence, confirmations, and cautions in interpreting the patterns of results in the full data set. The high-exposure data set provided more control over variables that have a substantial impact on the testing effect (initial test performance and feedback), and in some cases, covary with other moderators in the existing literature (e.g., studies employing cued recall initial tests more frequently employ feedback than those utilizing free recall). As such, discussion of the full data set results are supplemented with those of the high-exposure data set when pertinent, together providing a means to more accurately interpret moderator analyses that are at risk for substantive interactions between moderators. For the interested reader, Appendix B reports the results of an additional moderator analysis that was applied to another constrained data set composed of only those studies with retention intervals of at least 1 day, and Appendix C provides descriptive contingency tables to examine the clustering of select moderators of theoretical interest in the high-exposure

[2] A random-effects meta-regression model with multiple predictors was initially considered; however, the nature of the available data presented some limitations. In particular, the most impactful moderator (initial test performance) was not reported for a substantial portion of studies in the meta-analysis, and thus its utility as a predictor in a meta-regression model necessitated a sizable reduction in the number of effect sizes available for analysis. Even so, the general patterns of results from such a model that incorporated a number of the most impactful moderators (e.g., initial test performance, retention interval, feedback, initial test type, final test type) was consistent with the results reported below from the random-effects model.

data set (feedback, initial and final test types, and initial–final test matching).

Results

With the full data set, the mean weighted effect size from the random-effects model, $g = 0.50$ (CI [0.42, 0.58]), was greater than 0, indicating a reliable testing effect ($Z = 12.38$, $p < .001$) when contrasted with a restudy control condition. There was a high degree of heterogeneity among the samples included in the analysis ($tau^2 = 0.21$, $Q = 1{,}009.42$, $p < .001$), with a substantial majority of the overall variation between studies resulting from heterogeneity ($I^2 = 84.35$). To further describe the data, Figure 1 displays a stem-and-leaf plot of the effect sizes included in the analysis. The mean unweighted effect size was $g = 0.54$, and the distribution of effect sizes was slightly negatively skewed (skew $= -0.08$). The median effect size was $g = 0.55$. The majority (81%) of effect sizes were positive (i.e., testing effects), with 18% negative (i.e., restudy condition advantages) and 1% equal to 0. In addition to the primary analysis, two sensitivity analyses were performed by modifying the imputed value of *r* (see Method section) from .5 (used in the main analyses) to .25 and .75. The data set was strongly robust to such modifications. All patterns of results and statistical test outcomes for the primary random-effects analysis, and analyses of moderator variables, remained identical to those of the main ($r = .5$) analysis.

In the high-exposure data set, the mean weighted effect size, $g = 0.66$ (CI [0.56, 0.75]), was reliably greater than 0 ($Z = 14.10$, $p < 001$). Between-study heterogeneity was high, although smaller than in the full data set ($tau^2 = 0.15$, $Q = 429.57$, $p < .001$, $I^2 = 78.81$). The mean unweighted effect size was $g = 0.71$. Nearly all effect sizes (93%) were positive (i.e., testing effects).

In order to explore the effect of study characteristics on the testing effect, moderator analyses are reported. Of the categorical moderators tested, two were primarily descriptive variables about studies and samples: publication status and sample source. The remaining 11 categorical and three continuous moderators described methodological characteristics of studies. A summary of the results from the moderator analyses are described below. A more thorough treatment of moderators relevant to the theoretical issues described in the introduction follows in the Discussion.

Categorical Moderator Analyses

Results from the categorical moderator analyses for the full data set are reported in Table 1, and for the high-exposure data set in Table 2. The statistical reliability of the testing effect for specific levels within moderator analyses can be determined by noting whether 0 is present within the 95% CI range, in which case the effect was not reliably different than 0. A significant Q_B statistic indicates that differences exist between at least some levels of a given moderator.

Heterogeneity was detected between levels in the *publication status* moderator analysis. Studies classified as published had a larger mean weighted effect size ($g = 0.58$, CI [0.49, 0.67]) than those unpublished ($g = 0.25$, CI [0.10, 0.41]), although both were reliably greater than 0. However, caution is recommended in interpreting this difference, as no difference between published and unpublished effect sizes was present in the high-exposure data

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

Figure 1. Stem-and-leaf plot of effect sizes included in the meta-analysis.

Table 1
Full Data Set Categorical Moderator Analyses

Moderator	g	95% CI *LL*	95% CI *UL*	Q_B	k
Publication status				13.04**	
Published	0.58	0.49	0.67		122
Unpublished	0.25	0.10	0.41		37
Sample source				0.83	
College	0.49	0.40	0.58		136
Other	0.59	0.40	0.77		23
Design				5.22*	
Between	0.69	0.48	0.89		53
Within	0.43	0.35	0.52		106
Stimulus type				10.33*	
Prose	0.58	0.34	0.82		23
Paired associates	0.59	0.49	0.70		71
Single words	0.39	0.24	0.53		58
Other	0.27	0.06	0.48		7
Stimulus interrelation				1.82	
Prose	0.58	0.34	0.82		23
Categorical	0.48	0.22	0.73		20
No relation	0.50	0.41	0.60		111
Other	0.31	−0.01	0.63		5
List blocking				2.44	
Mixed	0.49	0.37	0.62		42
Blocked	0.46	0.34	0.57		87
Other	0.66	0.44	0.87		30
Initial test type				13.84**	
Cued recall	0.61	0.52	0.69		104
Free recall	0.29	0.07	0.52		36
Recognition	0.29	0.10	0.47		19
Retention interval (categorical)				11.65**	
<1 day	0.41	0.31	0.51		103
≥1 day	0.69	0.56	0.81		56
Initial test cue–target relationship				14.72**	
Same (recognition)	0.29	0.10	0.47		19
Nonsemantic	0.54	0.42	0.66		48
Semantic unrelated	0.67	0.41	0.94		13
Semantic related	0.66	0.51	0.82		43
None (free recall)	0.29	0.07	0.52		36
Final test type				7.38*	
Cued recall	0.57	0.46	0.68		71
Free recall	0.49	0.34	0.63		67
Recognition	0.31	0.15	0.46		21
Initial–final test match				1.00	
Different	0.58	0.45	0.71		56
Same	0.46	0.36	0.56		103
Feedback				18.72**	
No	0.39	0.29	0.49		107
Yes	0.73	0.61	0.86		52
Retrievability and reexposure				31.88**	
No feedback ≤50%	0.03	−0.21	0.27		17
No feedback 51%–75%	0.29	0.09	0.49		31
No feedback >75%	0.56	0.42	0.70		40
Feedback	0.73	0.61	0.86		52
Unknown and no feedback	0.48	0.24	0.71		19

Note. g = mean weighted effect size; CI = confidence interval; *LL* = lower limit; *UL* = upper limit; k = number of effect sizes.
* Q_B test for heterogeneity between levels of a moderator was significant at $p < .05$. ** Q_B test for heterogeneity between levels of a moderator was significant at $p < .01$.

set (g = 0.66 and 0.60, respectively). A closer look at the data suggests that unpublished reports typically had lower initial test performance and no feedback. The issue of publication bias will be returned to later in the Discussion.

In the sample source moderator analysis, effect sizes were homogeneous between levels representing studies utilizing college-enrolled participant samples and those utilizing samples drawn from other, noncollege or unspecified populations. As a result, it appears that the efficacy of testing is not dependent on the college student populations from which the vast majority of studies derive samples. Although a more fine-grained analysis of specific noncollege samples is desirable, there are not an

Table 2
High-Exposure Data Set Categorical Moderator Analyses

Moderator	g	95% CI *LL*	95% CI *UL*	Q_B	k
Publication status				0.42	
Published	0.67	0.56	0.77		79
Unpublished	0.60	0.42	0.78		13
Sample source				0.16	
College	0.66	0.55	0.77		71
Other	0.62	0.46	0.79		21
Design				13.61**	
Between	0.97	0.77	1.17		27
Within	0.55	0.46	0.64		65
Stimulus type				14.55**	
Prose	0.73	0.47	0.99		13
Paired associates	0.69	0.56	0.82		47
Single words	0.64	0.46	0.82		27
Other	0.33	0.18	0.48		5
Stimulus interrelation				4.44	
Prose	0.73	0.47	0.99		13
Categorical	0.56	0.26	0.86		13
No relation	0.67	0.56	0.78		63
Other	0.46	0.26	0.65		3
List blocking				1.12	
Mixed	0.64	0.51	0.77		26
Blocked	0.63	0.49	0.77		50
Other	0.77	0.53	1.01		16
Initial test type				14.14**	
Cued recall	0.72	0.61	0.83		65
Free recall	0.81	0.45	1.18		10
Recognition	0.36	0.19	0.52		17
Retention interval (categorical)				4.14*	
<1 day	0.58	0.47	0.70		57
≥1 day	0.78	0.63	0.94		35
Initial test cue–target relationship				15.85**	
Same (recognition)	0.36	0.19	0.52		17
Nonsemantic	0.61	0.48	0.74		28
Semantic unrelated	0.74	0.40	1.08		11
Semantic related	0.83	0.64	1.03		26
None (free recall)	0.81	0.45	1.18		10
Final test type				22.86**	
Cued recall	0.70	0.58	0.83		43
Free recall	0.79	0.61	0.97		32
Recognition	0.32	0.18	0.46		17
Initial–final test match				0.16	
Different	0.68	0.54	0.82		39
Same	0.64	0.52	0.76		53
Feedback				3.38	
No	0.56	0.42	0.70		40
Yes	0.73	0.61	0.86		52
Retrievability and reexposure				3.38	
No feedback >75%	0.56	0.42	0.70		40
Feedback	0.73	0.61	0.86		52

Note. g = mean weighted effect size; CI = confidence interval; *LL* = lower limit; *UL* = upper limit; k = number of effect sizes.
* Q_B test for heterogeneity between levels of a moderator was significant at $p < .05$. ** Q_B test for heterogeneity between levels of a moderator was significant at $p < .01$.

adequate number of studies from any particular noncollege demographic to allow for such an analysis. Preliminary evidence does, however, show reliable testing effects in older adults and children, in addition to in studies utilizing more diverse Internet sampling.

The moderator analysis of design yielded heterogeneity between levels, with those studies utilizing a between-participants design yielding larger effect sizes (g = 0.69, CI [0.48, 0.89]) than those studies utilizing a within-participant design (g = 0.43, CI [0.35, 0.52]).

The moderator analysis of stimulus type found heterogeneity between levels, with paired associates (g = 0.59, CI [0.49, 0.70]) and prose (g = 0.58, CI [0.34, 0.82]) showing the numerically largest effect size estimates. Single words (g = 0.33, CI [0.16, 0.50]) and other ($g = 0.27$, CI [0.06, 0.48]) stimulus types yielded numerically smaller but reliable effects. Note that in the high-exposure data set, single words produced similarly large effect sizes as paired associates and prose (see Table 2).

The moderator analysis of stimulus interrelation did not yield significant heterogeneity. This result suggests that the semantic relationships between materials, or lack thereof, did not have a reliable impact on the testing effect.

The moderator analysis of list blocking did not detect heterogeneity between levels. Of those studies utilizing list-learning procedures, similar magnitude, reliable testing effects were found in mixed ($g = 0.49$, CI [0.37,0.62]) and blocked ($g = 0.46$, CI [0.34, 0.57]) list designs.

The moderator analysis of initial test type yielded substantial heterogeneity between levels. Cued recall ($g = 0.61$, CI [0.52, 0.69]) yielded the largest testing effects, whereas free recall ($g = 0.29$, CI [0.07, 0.52]) and recognition ($g = 0.29$, CI [0.10, 0.47]) testing yielded more modest but reliable effects. Although cued recall resulted in a larger effect size than free recall in the full data set, this result should be interpreted cautiously given that cued recall testing was associated with more frequent use of feedback (44% of effect sizes vs. 11% following free recall) and somewhat higher average initial test performance (65% vs. 59% in cued recall and free recall, respectively), two factors that are strongly associated with the magnitude of the testing effect. Results from the high-exposure data set help to control for these discrepancies and, as such, show similar magnitude cued recall ($g = 0.72$, CI [0.61, 0.83]) and free recall ($g = 0.81$, CI [0.45, 1.18]) testing effects, both of which yielded significantly larger effects than recognition ($g = 0.36$, CI [0.19, 0.52]; ps $< .001$ and .03, respectively).

The categorical moderator analysis of retention interval yielded significant heterogeneity between levels. Studies with retention intervals of at least 1 day were associated with larger testing effects ($g = 0.69$, CI [0.56, 0.81]) than those with retention intervals less than 1 day ($g = 0.41$, CI [0.31, 0.51]).

The moderator analysis of initial test cue–target relationship found heterogeneity between levels; however, this largely resulted from the inclusion of the same and none levels in the analysis, which contained studies utilizing initial recognition and free recall testing, respectively. A reanalysis in which only the initial cued recall testing groups were included did not yield significant heterogeneity in the full data set ($Q_B = 1.83$, $p = .40$) or high-exposure data set ($Q_B = 4.15$, $p = .13$). Note, however, that numerical trends in the results map onto the general predictions of the elaborative retrieval view, which specifies that materials allowing for semantic relational processing should benefit the most from testing. These relevant data are considered more thoroughly in the Discussion.

The moderator analysis of final test type found heterogeneity between levels, with cued recall ($g = 0.57$, CI [0.46, 0.68]) producing a slightly higher point estimate than free recall ($g = 0.49$, CI [0.34, 0.63]) final testing and, in turn, a larger estimate than recognition final testing ($g = 0.31$, CI [0.15, 0.46]). All types of final tests led to statistically reliable testing effects. Note that feedback was employed more frequently in studies utilizing cued recall (46% of effect sizes) than free recall (18% of effect sizes) final tests. As such, the high-exposure data set likely provides more accurate estimates, with cued recall ($g = 0.70$, CI [0.58, 0.83]) and free recall ($g = 0.79$, CI [0.61, 0.97]) yielding significantly larger testing effects than recognition ($g = 0.32$, CI [0.18, 0.46]) final tests (ps $< .001$).

The moderator analysis of initial–final test match did not detect significant heterogeneity between levels. Both matching and mismatching testing formats yielded reliable testing effects.

The moderator analysis of feedback yielded significant heterogeneity between levels, with studies providing feedback (i.e., a re-presentation of originally studied information at least once after a testing opportunity) yielding larger testing effects ($g = 0.73$, CI [0.61, 0.86]) than studies not providing feedback ($g = 0.39$, CI [0.29, 0.49]). Despite the feedback advantage, studies not providing feedback still reliably produced testing effects, indicating that even when total exposure time to information is biased against the test condition (as tests are rarely performed with perfect accuracy), a benefit of retrieval over restudy can still reliably emerge. An additional analysis was conducted on only those studies providing feedback, partitioned according to whether the feedback was presented immediately following the retrieval attempt or after a delay. A difference was detected ($p = .02$), with delayed feedback ($g = 1.38$, $k = 6$) yielding a larger effect than immediate feedback ($g = 0.66$, $k = 46$). This result should be considered with caution given the relatively low number of observations in the delayed feedback condition and the fact that retention interval was not explicitly controlled for, as it can often be confounded with feedback delay (see Metcalfe, Kornell, & Finn, 2009; T. A. Smith & Kimball, 2010). Even so, the result does coincide with existing research demonstrating a benefit of delaying feedback following testing (Butler, Karpicke, & Roediger, 2007).

The moderator analysis of retrievability and reexposure yielded heterogeneity between levels. Studies not providing feedback yielded a positive relationship between initial retrieval success and the magnitude of the testing effect, such that initial test performance less than or equal to 50% did not produce a reliable testing effect ($g = 0.03$, CI [-0.21, 0.27], $p = .79$), whereas reliable testing advantages were found following 51%–75% initial retrieval success ($g = 0.29$, CI [0.08, 0.49]) and greater than 75% initial retrieval success ($g = 0.56$, CI [0.42, 0.70]). Studies with feedback, regardless of retrieval success, yielded the numerically largest testing effects ($g = 0.73$, CI [0.61, 0.86]). The group of studies that did not provide feedback nor report or record initial test performance yielded an effect ($g = 0.44$, CI [0.21, 0.67]) resembling the point estimate of the primary, full data set random-effect analysis.

Continuous Moderator Analyses

Results from the continuous moderator analyses are reported in Table 3 for the full data set and Table 4 for the high-exposure data set. The meta-regression models employed utilized the method of moments, with the results indicating the slope estimates, 95% CIs, and statistical reliability against the null hypothesis. The values reported can be interpreted in a way similar to a standard regression analysis. A positive slope indicates that effect sizes increase along with the variable under consideration (the magnitude of the slope indicates the steepness of the increase) according to the best fit model. The 95% CIs provide an indication of the reliability of the slope estimate from the meta-regression model. When 0 is not included within the interval, the association between effect size and the moderator was statistically reliable. The reported Z values

Table 3
Results From Full Data Set Continuous Moderator Meta-Regression Models

Moderator	Slope point estimate	*SE*	95% CI *LL*	95% CI *UL*	*Z*
Number of initial tests	0.01031	0.03374	−0.05581	0.07644	0.31
Initial test lag	0.00015	0.00016	−0.00017	0.00046	0.91
Log retention interval	0.08304	0.02651	0.03108	0.13500	3.13*

Note. CI = confidence interval; *LL* = lower limit; *UL* = upper limit.
* $p < .05$.

indicate significance tests of the slope against the null (i.e., no relationship between predictor and effect sizes).

In both data sets, only the continuous retention interval analysis was significant. As in the corresponding categorical moderator analysis, the magnitude of the testing effect appears to grow with the duration of the retention interval.

Discussion

Using meta-analysis, the present study demonstrates the reliability of the testing effect. The estimated mean weighted testing effect generated by the random-effects model was positive (g = 0.50, CI [0.42, 0.58]) and statistically reliable ($p < .001$). I first discuss the results in the context of broad theoretical characterizations of the testing effect. Next, results in relation to theoretically relevant boundary conditions of the testing effect, as described in the Introduction, are discussed. In addition, more thorough consideration is given to the bifurcation framework (Kornell et al., 2011) for characterizing the testing effect literature. Last, I note limitations of the present study, along with conclusions and recommendations for future investigations of the testing effect.

Broad Theoretical Implications

As described in the introduction, the primary purpose of the present meta-analysis was to evaluate existing theories of the testing effect and inform future theory building. For purposes of completeness, first note that all effect sizes were derived from studies utilizing a restudy control condition, and therefore the presence of a reliable testing effect rules out the historical explanation of increased test item exposure as the source of the testing effect. In fact, the reported testing effect estimates are conservative, in that exposure to material was often biased in favor of the control (restudy) conditions in which all information received full reexposure. With that noted, I next consider theoretical accounts of testing effect, including TAP theory and retrieval effort theories, in light of the present results.

TAP theory. TAP theory, as applied to the testing effect, specifies that testing effects may reflect the high degree of similarity in cognitive processes utilized during learning and assessment when testing is employed during learning (see Roediger & Karpicke, 2006a). Although this theory has been called into question as the sole explanation of the testing effect (e.g., Carpenter & DeLosh, 2006; Kang et al., 2007), one possibility is that the match in processing from learning to assessment may complement other mechanisms that are at play in the testing effect. The present study showed, however, that initial–final test match did not yield a reliable increase in the magnitude of the testing effect, nor was there a suggestion of a trend. This pattern was present in both the full data set and high-exposure data set. Thus, in light of the present results, it seems that TAP theory does not provide a viable explanation of the testing effect.

Note that Roediger and Karpicke (2006a) provide a possible means to reconcile the present results with TAP theory by suggesting that recall tests, more so than recognition tests, may generally induce the more effortful processing that is necessary to perform well on an assessment after a retention interval. Indeed, this may be a possibility given the present results. However, as noted by Roediger and Karpicke, this is not an a priori prediction of TAP theory.

Retrieval effort theories. Retrieval effort theories, considered as a class, typically postulate that the testing effect results from the effort, intensity, or depth of processing induced during an initial test (see, e.g., Bjork, 1975; Bjork & Bjork, 1992; Glover, 1989; Pyc & Rawson, 2009). As described in the Introduction, retrieval effort theories are not always specific as to what it is about retrieval that strengthens memory. Nonetheless, the results of the meta-analysis are somewhat consistent with the major predictions deriving from such theories. A primary prediction of

Table 4
Results From High-Exposure Data Set Continuous Moderator Meta-Regression Models

Moderator	Slope point estimate	*SE*	95% CI *LL*	95% CI *UL*	*Z*
Number of initial tests	0.00430	0.03469	−0.06369	0.07229	0.12
Initial test lag	0.00009	0.00017	−0.00025	0.00042	0.51
Log retention interval	0.06238	0.02942	0.00472	0.12005	2.12*

Note. CI = confidence interval; *LL* = lower limit; *UL* = upper limit.
* $p < .05$.

retrieval effort theories, collectively, is that the type of initial test should dictate the magnitude of the testing effect. More effortful or difficult tests (i.e., recall more than recognition, typically) should yield larger testing effects. The results of the meta-analysis largely support this prediction. Given the clustering of feedback and initial test performance with initial test type, the high-exposure data set provides the most appropriate analysis of initial test type. The test type that presumably places the least demands on retrieval, recognition, yielded the smallest effect size ($g = 0.36$), whereas cued recall and free recall led to substantially larger effects ($g = 0.72$ and 0.82, respectively). In addition, free recall, which is presumably more difficult than cued recall, yielded a numerically higher point estimate in the high-exposure data set, although the difference was not significant. Even so, the results of the initial test type moderator analysis fit with retrieval effort theories if one considers recall and recognition to reflect, or differentially induce, qualitatively different processes. Given this characterization, the aspects of a test necessary to yield a large mnemonic advantage, beyond that gained through restudy, appear to derive heavily from retrieval processes, more so than recognition processes. Recognition testing did, however, yield a reliable testing effect, coinciding with the results of a number of classroom studies of the testing effect (e.g., Roediger et al., 2011). Recent work also shows that recognition tasks can be crafted in such a way to induce extensive retrieval processes (Little et al., 2012) and can thereby provide even more effective practical utility in promoting learning. Moreover, even though the present meta-analysis focuses on direct benefits of testing, tests of any type can benefit memory through indirect means. For example, tests can enhance a learner's metacognitive awareness of content that is or is not well learned, thereby allowing for more efficient subsequent study. Thus, the application of any type of testing, including recognition testing, is likely to produce a more robust retention benefit in a nonlaboratory setting than is suggested by the present meta-analysis of laboratory studies.

An additional analysis was carried out to further test retrieval effort theories. In addition to serving as a proxy for item exposure, initial test performance can be used to estimate retrieval difficulty (i.e., low initial test performance suggests a difficult initial test). As such, low initial test performance, if not confounded by low test condition exposure, could plausibly be predicted to correlate with larger testing effects based on the assumption that the testing procedure demanded more effortful retrieval. Controlling for item exposure by using only studies that provided feedback, an analysis was conducted treating initial test performance (drawing from the first initial test when data were available) as a moderator. Heterogeneity was detected ($p < .01$), such that studies utilizing feedback with low (<50%) initial test performance yielded numerically larger effects ($g = 0.99$) than those with moderate (51%–75%; $g = 0.68$) and high (>75%; $g = 0.40$) initial test performance, thereby providing an additional result in support of retrieval effort theories. Even so, caution is advised in interpreting this result as robust, given the constrained data set.

A second prediction drawing from some theories classified under the umbrella of retrieval effort is that the number of retrievals should be positively related to the magnitude of the testing effect (Glover, 1989). Contrary to this prediction, no relationship was found between number of initial tests and the magnitude of the testing effect: A single test yielded a similar magnitude effect as compared to multiple tests. This result may seem at odds with individual studies showing that the magnitude of the testing effect is positively related to the total number of retrievals (e.g., Karpicke & Roediger, 2008; McDermott, 2006; Roediger & Karpicke, 2006b; Vaughn & Rawson, 2011; Zaromb & Roediger, 2010). As such, the result should be interpreted with caution, and likely results from the conservative orientation of the present meta-analysis (i.e., considering only those studies with restudy control conditions). As a result, those studies with repeated initial tests also presented their restudy control condition information repeatedly. This may serve at least to mitigate any relative benefit of repeated testing. Glover (1989) did not specify a mechanism by which repeated retrievals should boost performance, although one may suspect that processing that may be induced by an initial test (e.g., semantic elaboration) is largely sufficient (if only for limited time), with additional, temporally close tests providing limited additional benefit. Indeed, although Pyc and Rawson (2009) found that final test performance increased with the number of successful retrievals, the returns diminished with each consecutive retrieval. Furthermore, not all studies utilizing repeated testing procedures provided feedback following an initial test, thus limiting the extent to which repeated testing allows one to retain previously unretrievable information. In contrast, given repeated study, all material is reexposed during each restudy opportunity. Given that the testing effect seems to grow with the retention interval, the potential benefit of repeated testing, especially if strengthening only a subset of retrievable items, is likely overshadowed by repeated restudy (strengthening all items), especially so at short retention intervals. It may also be the case that the high degree of between-study heterogeneity introduced enough variability to conceal a subtle effect. Regardless, a key implication of the present result is that even a single test can provide a substantial mnemonic benefit. Placed in the context of the larger literature, repeated testing is likely beneficial, even if returns diminish across trials.

Overall, retrieval effort theories largely gathered support from the present meta-analysis. Recall tasks, more so than recognition tasks, produced large testing effects, though recognition tasks alone are sufficient to induce reliable benefits. The results do not support TAP theory, in that the match between initial and final test formats does not appear to be an important factor in the testing effect. Results that pertain to the elaborative retrieval hypothesis, the bifurcation model, and other theoretically relevant characteristics of the testing effect are described below.

Boundary Conditions of the Testing Effect

A second contribution of the present meta-analysis is to examine the reliability of potential moderators and help establish the boundary conditions of the testing effect across the literature. Assessing the reliability of such factors has direct implications for both the interpretation of existing data and future theoretical development. Three factors are discussed: the influence of experimental design, the impact of the retention interval, and the durability of testing effects across materials.

Aspects of experimental design can have a robust impact on the emergence and size of numerous memory phenomena (see, e.g., McDaniel & Bugg, 2008). However, it is unclear whether such design factors reliably impact the testing effect in a comparable way. The generation effect, as a retrieval-based memory phenomenon, provides an obvious comparison by which to contrast the testing effect. The generation effect refers to the finding that during learning, generating stimuli can boost retention more so than simply studying information. Testing and generation are frequently regarded as synonymous and interchangeable. Indeed, as Karpicke and Zaromb (2010) describe the current state of the literature, "there is currently no well-developed empirical or theoretical basis to distinguish the effects" (p. 228). To this end, two moderator analyses of interest were examined: design and list blocking. Results indicated that design influenced the magnitude of the testing effect. However, whereas generation effects appear larger with within-participant designs (Berstsch et al., 2007), the opposite was found in regard to the testing effect, with between-participants designs leading to larger effects ($g = 0.69$, compared with $g = 0.43$ for within-participant designs).

Additionally, although list blocking impacts the magnitude of the generation effect (Berstsch et al., 2007), it was not reliably associated with the magnitude of the testing effect. Of those studies utilizing list-learning procedures, mixed lists produced similar magnitude, reliable testing effects ($g = 0.49$) as blocked lists ($g = 0.46$). This pattern of results is in contrast to the generation effect described in a meta-analysis of that literature (i.e., see the effect sizes grouped by list blocking condition reported in Berstsch et al., 2007). However, the theoretical account outlined in the introduction that suggests generation encourages item-specific encoding at the expense of serial order encoding (see McDaniel & Bugg, 2008; Nairne et al., 1991) specifically applies to free recall final testing, in which relational information is of particular use in guiding recall in the absence of other cues. As such, a supplemental analysis was run on a restricted data set including only those studies utilizing free recall final tests. Heterogeneity was not detected ($p = .25$), though studies utilizing mixed lists did lead to a numerical advantage ($g = 0.59$, CI [0.40, 0.78], $p < .001$, $k = 19$) over studies utilizing blocked lists ($g = 0.42$, CI [0.19, 0.64], $p < .001$, $k = 40$). In both cases, statistically reliable testing effects were obtained, suggesting a distinction from the pattern of results indicated in the generation effect literature, where blocked lists typically lead to null or negative generation effects in free recall (e.g., Serra & Nairne, 1993). Note that the impact of list blocking does not directly speak to the applicability of the item-order account to the testing effect, as it does not directly assess order memory. Even so, the robustness of testing across list designs (see also Rowland et al., 2014) suggests that somewhat different dynamics are at play in testing paradigms than in generation, perhaps resulting from the availability of an episodic learning context associated with retrieved information in the case of the testing effect (see, e.g., Rowland & DeLosh, 2014a).

A second boundary condition that has drawn much attention concerns the relationship between the testing effect and the retention interval. Although the testing effect reliably emerges following long retention intervals, mixed support has been found for short intervals, with some reports showing that restudying information during learning is more effective than testing at promoting short-term retention (e.g., Roediger & Karpicke, 2006b; Toppino & Cohen, 2009; Wheeler et al., 2003). The present meta-analysis does show that testing effects are larger in magnitude following longer retention intervals, confirming the pattern of results often observed in the literature. However, an additional key finding is that testing reliably benefits retention compared with restudy even at short intervals. The observation of a reliable short-term testing effect is potentially problematic for some explanations that have been advanced to explain the testing by retention interval interaction, which predict null or negative testing effects at short intervals (e.g., the idea that restudy, compared with testing, may benefit initial learning at the expense of more rapid forgetting; see, e.g., Wheeler et al., 2003). Individual studies that show a restudy benefit at short retention intervals may in part reflect low initial test performance (and thus lack of reexposure) for test condition items (e.g., Wheeler et al., 2003), given that other studies with short retention intervals and high initial test performance (e.g., Carpenter, 2009; Carpenter & DeLosh, 2006; Kuo & Hirshman, 1996; Rowland & DeLosh, 2014b) or feedback (e.g., Carrier & Pashler, 1992; Kang, 2010) show significant testing effects. As such, findings of null or negative effects of testing at short intervals may reflect factors that dictate initial testing performance or test item exposure, rather than, or in addition to, other possible mechanisms for the interaction.

A third point of interest concerns the impact of materials employed to investigate the testing effect. Of particular relevance is that the testing effect does not seem to be dependent on the learning of a specific type of material. Across the literature, both verbal and nonverbal materials yield reliable testing effects. According to the elaborative retrieval hypothesis of the testing effect, retrieval induces more elaborative processing than restudy, thus increasing the likelihood of subsequent retrieval. The evidence for this view, however, comes entirely from studies using verbal materials. It is not clear how memory for pictorial or spatial information would allow for elaboration, at least of the same form, and given the existing specification of the elaborative retrieval view, a testing effect would not necessarily be predicted when nonverbal materials are used. Yet, Kang (2010) had participants who were unfamiliar with the Chinese language learn Chinese characters, each paired with an English cue word. On a retention test, memory for the characters was better for those participants who had previously recalled the characters (by drawing them given the English cue word) than those who had studied the characters an equivalent amount of time. Similar results have been found for the retention of spatial information in the learning of maps (Carpenter & Pashler, 2007; Rohrer et al., 2010), with testing proving to be a more effective learning strategy than study. In addition, Carpenter and Kelly (2012) demonstrated a testing effect for the learning of spatial relationships between items in three-dimensional space. In light of these studies, it seems that theories ascribing verbally based mechanisms to the testing effect are not readily or unambiguously applicable to certain demonstrations of the testing effect reported in the literature. To account for the full range of testing effects that have been reported, a combination of factors may be needed. Following from this, one means to assess existing theoretical accounts is to consider their ability to predict the magnitude of the testing effect given certain types of materials and learning conditions.

Concerning the types of materials used (paired associates, single words, prose, and other, miscellaneous materials), the somewhat better retention for studies using paired associates and prose relative to other materials found in this meta-analysis is partially consistent with one recently proposed theoretical mechanism. Pyc and Rawson (2010) introduced the mediator effectiveness hypothesis, which, consistent with the elaborative retrieval hypothesis, suggests that mediating information (i.e., information that links, in some way, a cue to a target) can be more effectively utilized by testing compared with study procedures. Carpenter (2011) elaborated on this view by demonstrating that engaging in a test can activate mediating information that provides a semantic link between cue and target, later serving to facilitate target access. Following directly from this view, paired associates may have an opportunity to benefit from the generation and utilization of mediating information, whereas other types of materials may not. Granted the assumption that such elaborative semantic processing may occur with other verbal materials (e.g., prose), the results of the stimulus type moderator analysis are compatible with such theories of the testing effect. Note, however, that Pyc and Rawson explicitly state that enhancing the utility of mediators may best be considered as a contributor rather than a complete mechanistic explanation of the testing effect.

The results of the initial test cue–target relationship moderator analysis allow for a more fine-grained examination of the impact of semantic information between cue and target during testing. Cue–target pairs that allow for semantic relational processing may benefit from testing to a greater degree than materials that do not encourage such processing, perhaps through the generation of semantic mediating information (Carpenter, 2011). Furthermore, those materials that do allow for cue–target semantic relational processing may differentially benefit depending on the degree of existing semantic relatedness between the cue and the target themselves. That is, less semantically associated cues and targets may allow for more thorough elaboration prior to target retrieval (Carpenter, 2009). The results of the meta-analysis do not provide compelling support for such theories of the testing effect. However, although the relevant levels of the initial test cue–target relationship moderator analysis (i.e., nonsemantic, semantic unrelated, and semantic related) did not reliably differ from each other, the numerical pattern of results fits partially with the predictions outlined above. Whereas cue–target pairs not readily allowing for relational semantic processing did yield reliable testing effects ($g = 0.54$), pairs with semantically related cues and targets yielded numerically larger effects ($g = 0.66$), as did materials employing semantically unrelated cue and target pairs ($g = 0.67$). A similar, though more pronounced pattern was evident in the high-exposure data set (see Table 2). Note that the degree of semantic relatedness between cues and targets did not influence the magnitude of the testing effect ($g = 0.66$ vs. 0.67), in contrast with a prediction of the elaborative retrieval view (see Carpenter, 2009). However, given the existing, limited body of literature on the topic, a more detailed analysis may be better left to individual studies if the effect is modest. The more general prediction of larger effects for semantic than nonsemantic cue–target pairs is more suitable for assessment through meta-analysis at this time. As such, the present results do not clearly endorse semantic elaboration as the major contributing mechanism to the testing effect, though with the caveat that the high degree of heterogeneity and limited studies in the area may preclude any strong confirming or disconfirming conclusions regarding such theories.

A third factor concerning the impact of materials on the testing effect concerns the relationship among to-be-learned information. Some investigators have theorized that testing may effectively promote the processing of gist or thematic information underlying to-be-learned information (e.g., Verkoeijen et al., 2012), or similarly, may promote semantic organization of information in memory (e.g., Congleton & Rajaram, 2012). One potential consequence of this is that the presence of common, clearly noticeable semantic features across materials may, in some cases, be redundant with the processing engaged in during testing, and thus negatively impact the magnitude of the testing effect (see, e.g., Delaney et al., 2010). Alternatively, conceptual commonalities across to-be-learned materials could potentially be exploited through testing, thus increasing the magnitude of the effect (see, e.g., Congleton & Rajaram, 2012). The results from the stimulus interrelation moderator analysis failed to find significant heterogeneity across groups of studies using materials that contained either no underlying theme (e.g., unrelated word lists), an unstructured semantic theme (e.g., categorized lists), or integrated, conceptual relations (e.g., prose), with similar magnitude effect sizes found in each group.

One potentially mitigating factor for this result is that those studies classified as having semantically themed materials included the use of both single-theme lists (e.g., DRM lists; e.g., McConnell & Hunt, 2007; Verkoeijen et al., 2012) and lists representing multiple themes (e.g., Congleton & Rajaram, 2012; Zaromb & Roediger, 2010). A plausible prediction is that materials representing multiple categories of information may benefit from enhanced organization (Zaromb & Roediger, 2010) and gist processing (Verkoeijen et al., 2012) resulting from testing, especially when the categorical nature is not overly explicit. A follow-up analysis subdividing the semantically themed group according to the presence of single versus multiple categories represented per list did not detect heterogeneity ($p = .84$). However, given the limited number of existing studies using categorized materials ($k = 20$), these analyses should be viewed as exploratory. Furthermore, the impact of categorization may interact with additional moderators such as the cues available at final testing, the retention interval, and potentially other fine-grained design factors (for discussion of such possibilities, see Congleton & Rajaram, 2012; Verkoeijen et al., 2012; see also Peterson & Mulligan, 2013). Presently, the impact of relational and organizational processing mechanisms during testing are not clear, and the topic would benefit from additional research.

In sum, the impact of to-be-learned materials on the testing effect appears to be limited, although theoretically consistent trends were found in effect sizes, such that paired associates and prose materials, both allowing for relational processing, may benefit somewhat more from testing than other materials. However, robust testing effects are found across highly variable materials, and thus there are likely multiple mechanisms at play that ultimately yield a test-induced benefit to memory.

A Bifurcation Framework for the Testing Effect

The recently developed bifurcation model of the testing effect (Halamish & Bjork, 2011; Kornell et al., 2011) can provide a useful framework to conceptualize the patterns of results in the testing effect literature. According to the bifurcation model, test condition and study condition items can each be represented by independent distributions along a continuum of memory strength. During initial study, both sets of items are treated equally, and thus both item distributions receive similar increments in memory strength. However, at the time of initial test or restudy, the item distributions begin to disperse. Because all study condition items are granted a second exposure, the entire study distribution gains some degree of memory strength. Test condition items, however, are rarely recalled to perfection during initial testing. For those items in the test distribution above the initial test threshold, retrieval is successful, granting a large increment to memory strength. However, those items below threshold are not retrieved and thus do not benefit from the failed test (assuming no feedback is provided). This variable treatment of test condition items, which is dependent on retrieval success, is modeled as a bifurcation in the test distribution. That is, the subset of the test distribution that is successfully retrieved increases in memory strength (and importantly, to a greater degree than the study distribution following restudy), whereas the unsuccessfully retrieved test items (i.e., the portion of the test distribution below threshold) remain stationary with regard to memory strength and do not benefit from the test. At assessment, retrieval will only be successful for items represented by the portions of either distribution that cross above the final test memory strength threshold, which in turn is determined by the difficulty of the final test (more difficult tests set higher thresholds for successful retrieval).

Although the bifurcation model is agnostic to any factors or mechanisms that may influence the degree of memory strength gained from initial retrieval (e.g., initial test type or stimulus characteristics), it does provide a useful means of framing the present results. Foremost, increasing the proportion of items successfully retrieved during an initial test should strengthen the testing effect by increasing the strength of a larger proportion of the test distribution, as was confirmed in this meta-analysis. Furthermore, the framework predicts that, all else constant, the testing effect should increase with final test difficulty. Or conversely, the relatively more modest enhancement in strength applied to a study distribution through restudy can overshadow a much larger strength increment applied to a smaller subset of the test distribution if the final test threshold is low enough to capture a large enough potion of the study distribution. This prediction was confirmed by the finding that recognition final tests (assumed to be easier tests, all else equal) appear to yield more modest ($g = 0.31$) testing effects compared with presumably more difficult cued recall ($g = 0.57$) and free recall ($g = 0.49$) tests in the full data set, and a more exaggerated pattern in the high-exposure data set ($g =$ 0.32, 0.70, and 0,79, respectively). The framework additionally predicts that final test difficulty can be increased by lengthening the retention interval, a prediction supported by the present findings. In addition, the framework can neatly account for the reliable testing effect found at short retention intervals (i.e., with presumably easier final tests) when initial test performance is sufficiently high or if feedback is provided.[3] Indeed, the magnitude of the continuous retention interval moderator slope was lower in the high-exposure data set compared with the full data set (i.e., retention interval had a smaller effect in the high-exposure data set), in accordance with the predictions of the model (Kornell et al., 2011).

Although the bifurcation model is generally consistent with the analyses reported here, it does not specify the mechanisms or conditions that would be expected to influence the magnitude of memory strength increment resulting from initial testing. As such, one way to consider the variables found to influence the testing effect (e.g., initial test type) is to frame them in the context of the bifurcation model. For instance, although initial recognition has produced somewhat inconsistent results in individual studies, it could be the case that the ability to detect a positive effect only emerges under sufficiently difficult final test conditions following high initial test performance. The framework may thus serve to be particularly useful in guiding future research on the testing effect, at the minimum by drawing attention to the conditions of the final assessment.

In sum, the results of the present meta-analysis are mostly consistent with the retrieval effort class of testing effect theories but provide limited support for the elaborative retrieval hypothesis in specifying a contributing mechanism. Additionally, the results suggest that the bifurcation model can provide a useful framework to interpret results and generate novel predictions. Support was not found for TAP theory as a major contributor to the testing effect.

Limitations and Future Directions

As with all meta-analyses, the present investigation has a number of limitations. A primary issue is that of publication bias (often referred to as the file drawer problem; Rosenthal, 1979). That is, because most academic journals rarely publish null results, the published literature may provide a skewed portrait of the phenomena of interest (i.e., null results are likely to gather dust in one's file drawer). Although there does not seem to be a universally appropriate method for assessing and correcting for publication bias, there are nonetheless a variety of commonly employed techniques (cf. Ferguson & Brannick, 2012; Rothstein & Bushman, 2012).

A direct means to reduce publication bias is to include unpublished studies in a meta-analysis. For the present meta-analysis, I included both published and unpublished studies that met inclusion criteria. An indication of possible publication bias was detected from the analysis of publication status in the full data set, such that published studies produced larger testing effects; unpublished studies produced smaller, though statistically reliable effects. Although there are techniques that are often used in meta-analyses to provide an indication as to the presence or extent of publication bias, no single procedure is globally preferred (Sutton, 2009). More important, many of the more popular techniques provide skewed to potentially misleading information, especially in the presence of substantial heterogeneity, as was the case in the present study (see, e.g., Ioannidis & Trikalinos, 2007; Terrin, Schmid, Lau, & Olkin, 2003). A more useful means for the present meta-analysis was to investigate whether the unpublished reports were biased in a manner that related to identified factors that impact the testing effect. This was the case, with effect sizes derived from unpublished studies ($k = 37$) appearing to cluster within two moderators that were found to substantially impact the magnitude of the testing effect: feedback (which was absent in 89% of effect sizes derived from unpublished studies) and relatively low initial test performance ($M = 59\%$ for effect sizes derived from those unpublished studies reporting the data). The high-exposure data set mitigated these confounds and, in the

[3] The bifurcation model does not specify the effect of feedback. However, given a conservative assumption that feedback provides no function beyond an additional exposure, it should at least reduce the degree of bifurcation evident in the test distribution (i.e., by strengthening the unsuccessfully retrieved items; see Kornell et al., 2011, for elaboration on this idea). As such, for present purposes, the inclusion of feedback should serve a function somewhat similar to increasing initial test performance.

publication status analysis, revealed that unpublished studies produced similarly sized testing effects as published studies. As a result, the unreliable testing effect found in the unpublished studies in the full data set was very likely attributable to the designs and procedures employed, thereby leading to a negatively biased effect estimate.

An additional limitation of meta-analysis concerns the potential for characteristics of included studies to covary with each other. Given that data sources in meta-analysis (i.e., existing studies) are fixed, and not under the control of the meta-analyst, common experimental design patterns in a given literature will naturally occur and must be taken into consideration. Indeed, this was evident in the present study. In order to mitigate the threat of moderator covariation skewing the results, analyses were conducted on a high-exposure data set along with the full data set. A primary purpose of the high-exposure data set was to help reduce the degree of variability that exists in the testing effect literature with regard to test condition item exposure, given that this factor was identified as particularly consequential. Even so, not all moderator clustering could be controlled for because of the nature of the testing effect literature, and as such, certain analyses were susceptible to significant moderator covariation. However, I suggest that cautious interpretation of the analyses from both the full and high-exposure data sets as appropriate, considered in tandem with established themes in the testing effect literature, can adequately mitigate the impact of misleading results in the present study. Furthermore, supplemental analyses on an additional restricted data set (long retention interval studies) are reported in Appendix B, providing additional analyses that partially control for a moderator of interest—retention interval—in the testing effect literature.

To maximize the informative value of future meta-analyses, it would be of great value to include full, detailed reporting of descriptive statistics and methodology in future publications concerning the testing effect. Both the calculation of effect sizes and the coding of detailed moderators can be done with greater accuracy when means, variance terms, and statistical test outcomes are reported with precision for all cells. Relevant to research on the testing effect specifically, many reports (studies contributing 27% of effect sizes) did not include data on initial test performance, or utilized protocols in which it was not gathered. Such data are important for both descriptive and theoretical reasons, and missing data limit the types and accuracy of possible analyses concerning the interaction between initial retrieval success and other potential moderators. The retrievability and reexposure moderator analysis demonstrates that testing effects are influenced, to a substantial degree, by successful initial retrieval. Furthermore, nonperfect initial test performance, if not taken into consideration, can act as a serious confound that can lead to misinterpretations of empirical results. Thus, one contribution of the present meta-analysis is to help bring attention to the issue of item exposure in testing effect research. Future investigations of the testing effect may benefit from acknowledging this confound, whether by inducing more consistent initial test performance, providing supplemental analyses on retrieved or unretrieved subsets of data, or simply considering the impact of initial test performance when interpreting data.

Conclusions

The findings of this meta-analysis make several key contributions to the literature. The results generally support the retrieval effort class of theories and the bifurcation model of the testing effect and suggest that semantic elaboration may potentially contribute to the testing effect but, at best, is not viable as a stand-alone mechanism of the effect. Further, the results of the meta-analysis are not consistent with TAP theory. Specifically, matching initial and final tests did not contribute any increase in the magnitude of the testing effect. Instead, initial recall tests produced larger magnitude testing effects than initial recognition tests. The importance of initial test type has implications for both theoretical explanations of the testing effect and the effective application of testing to everyday contexts. Analyses of final test type mirrored those of initial test type; recall final tests led to larger testing effects than recognition final tests. Taken together with the analyses of retention intervals (testing effects increase in magnitude with longer retention intervals), the results provide support for the bifurcation model of the testing effect. Numerical trends in the results suggested that design characteristics allowing for more substantial semantic processing and elaboration may lead to larger testing benefits on retention, consistent with the elaborative retrieval hypothesis (Carpenter, 2009, 2011; Carpenter & DeLosh, 2006). These trends were not statistically reliable in the case of the initial test cue–target relationship analysis, however, and thus enhanced semantic elaboration may only partially contribute to the testing effect or contribute to the testing effect in only circumscribed situations. Additional research is needed to identify other potentially important mechanisms that contribute to the testing effect, including episodic or other context-based mechanisms that have the potential to be uniquely exploited by testing (i.e., retrieving information from a specific past episode).

The present meta-analysis also quantitatively assessed the reliability of potential boundary conditions that may have implications for further theoretical characterization of the testing effect. The testing effect appears insensitive to manipulations of list blocking, potentially differentiating the effect from other memory phenomena (see, e.g., McDaniel & Bugg, 2008), including a similar retrieval-based effect: generation. In addition, results showed a positive relationship between the testing effect and the length of the retention interval. The retention interval analyses indicated that the magnitude of the testing effect increases over time, as has been reported in the literature. However, a reliable effect still obtained at short intervals on the order of minutes. Although certain circumstances produce a disordinal interaction between learning method (testing or restudy) and retention interval (see, e.g., Delaney et al., 2010; Roediger & Karpicke, 2006a), data from the literature suggest that testing can be beneficial at a wide variety of both short and long retention intervals, though perhaps to varying degrees. The results also draw attention to the importance of considering initial test performance when interpreting testing effect data—a variable demonstrated to be of substantial importance that is often not given due consideration in the testing effect literature.

Although not directly examined in the present meta-analysis, additional contributions to the testing effect beyond those of a semantic nature (e.g., semantic elaboration) may result from the processing of contextual information during retrieval opportunities. Testing promotes list differentiation (i.e., determining which of multiple lists a given item belongs to; Chan & McDermott, 2007; cf. Brewer, Marsh, Meeks, Clark-Foos, & Hicks, 2010) can reduce interference (Nunes & Weinstein, 2012; Szpunar, McDermott, & Roediger, 2008; Weinstein, McDermott, & Szpunar, 2011; see also Halamish & Bjork, 2011; Potts & Shanks, 2012), may promote encoding variability (McDaniel & Masson, 1985), and may increase access to past episodic or temporal contexts (Rowland & DeLosh, 2014a). These effects may all result from the enrichment, or strengthened association, of a memory trace with episodically linked contextual elements that can assist later retrievability. Given this possibility, the relative or joint contributions of semantic and contextual episodic information may be of interest to future theoretical investigations of the testing effect.

In conclusion, despite the robust nature of the testing effect, the underlying mechanisms that produce the effect remain elusive. Even so, recent theoretical developments have begun to apply greater specificity to existing characterizations of the effect. The present meta-analysis can help clarify a number of open questions in the literature and guide future theoretical development. I contend that the testing effect is likely to reflect multiple memory mechanisms, with the role of each dependent on the specific conditions involved. Future work may benefit from considering episodic or contextual contributions to memory that result from testing. A careful, thorough consideration of the factors that reliably influence the effectiveness of testing, as suggested by the present meta-analysis and other reports, will contribute to the development of a comprehensive theory of the testing effect.

References

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

Abbott, E. E. (1909). On the analysis of the factors of recall in the learning process. *Psychological Review: Monographs Supplements, 11,* 159–177. doi:10.1037/h0093018

Agarwal, P. K. (2012). Advances in cognitive psychology relevant to education: Introduction to the special issue. *Educational Psychology Review, 24,* 353–354. doi:10.1007/s10648-012-9212-0

Allen, G. A., Mahler, W. A., & Estes, W. K. (1969). Effects of recall tests on long-term retention of paired associates. *Journal of Verbal Learning and Verbal Behavior, 8,* 463–470. doi:10.1016/S0022-5371(69)80090-3

Anderson, J. R., & Bower, G. H. (1972). Recognition and retrieval processes in free recall. *Psychological Review, 79,* 97–123. doi:10.1037/h0033773

Anderson, M. C. (2003). Rethinking interference theory: Executive control and the mechanisms of forgetting. *Journal of Memory and Language, 49,* 415–445. doi:10.1016/j.jml.2003.08.006

Anderson, M. C., Bjork, R. A., & Bjork, E. L. (1994). Remembering can cause forgetting: Retrieval dynamics in long-term memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20,* 1063–1087. doi:10.1037/0278-7393.20.5.1063

Bangert-Drowns, R. L., Kulik, J. A., & Kulik, C. C. (1991). Effects of frequent classroom testing. *Journal of Educational Research, 85,* 89–99. doi:10.1080/00220671.1991.10702818

Bertsch, S., Pesta, B. J., Wiscott, R., & McDaniel, M. A. (2007). The generation effect: A meta-analytic review. *Memory & Cognition, 35,* 201–210. doi:10.3758/BF03193441

*Bishara, A. J., & Jacoby, L. L. (2008). Aging, spaced retrieval, and inflexible memory performance. *Psychonomic Bulletin & Review, 15,* 52–57. doi:10.3758/PBR.15.1.52

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

Bjork, R. A. (1988). Retrieval practice and the maintenance of knowledge. In M. M. Gruneberg, P. E. Morris, & R. N. Sykes (Eds.), *Practical aspects of memory: Current research and issues* (Vol. 1, pp. 396–401). New York, NY: Wiley.

Bjork, R. A., & Bjork, E. L. (1992). A new theory of disuse and an old theory of stimulus fluctuation. In A. Healy, S. Kosslyn, & R. Shiffrin (Eds.), *From learning processes to cognitive processes: Essays in honor of William K. Estes* (Vol. 2, pp. 35–67). Hillsdale, NJ: Erlbaum.

Borenstein, M. (2009). Effect sizes for continuous data. In H. Cooper, L. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis* (pp. 221–235). New York, NY: Russell Sage Foundation.

Bouwmeester, S., & Verkoeijen, P. P. J. L. (2011). Why do some children benefit more from testing than others? Gist trace processing to explain the testing effect. *Journal of Memory and Language, 65,* 32–41. doi:10.1016/j.jml.2011.02.005

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. doi:10.1080/09658211003702163

*Brewer, G. A., & Unsworth, N. (2012). Individual differences in the effects of retrieval from long-term memory. *Journal of Memory and Language, 66,* 407–415. doi:10.1016/j.jml.2011.12.009

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

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

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

Campbell, J., & Mayer, R. E. (2009). Questioning as an instructional method: Does it affect learning from lectures? *Applied Cognitive Psychology, 23,* 747–759. doi:10.1002/acp.1513

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

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

*Carpenter, S. K., & DeLosh, E. L. (2005). Application of the testing and spacing effects to name learning. *Applied Cognitive Psychology, 19,* 619–636. doi:10.1002/acp.1101

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

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

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

Carpenter, S. K., Pashler, H., & Cepeda, N. J. (2009). Using tests to enhance 8th grade students' retention of U.S. history facts. *Applied Cognitive Psychology, 23,* 760–771. doi:10.1002/acp.1507

*Carpenter, S. K., Pashler, H., & Vul, E. (2006). What types of learning are enhanced by a cued recall test? *Psychonomic Bulletin & Review, 13,* 826–830. doi:10.3758/BF03194004

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

*Carrier, M., & Pashler, H. (1992). The influence of retrieval on retention. *Memory & Cognition, 20,* 633–642. doi:10.3758/BF03202713

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

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

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. doi:10.1037/0278-7393.33.2.431

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

Congleton, A. R., & Rajaram, S. (2011). The influence of learning methods on collaboration: Prior repeated retrieval enhances retrieval organization, abolished collaborative inhibition, and promotes post-collaborative memory. *Journal of Experimental Psychology: General, 140,* 535–551. doi:10.1037/a0024308

*Congleton, A., & Rajaram, S. (2012). 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. doi:10.3758/s13421-011-0168-y

*Coppens, L. C., Verkoeijen, P. P. J. L., & Rikers, R. M. J. P. (2011). Learning Adinkra symbols: The effect of testing. *Journal of Cognitive Psychology, 23,* 351–357. doi:10.1080/20445911.2011.507188

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

*Cull, W. L. (2000). Untangling the benefits of multiple study opportunities and repeated testing for cued recall. *Applied Cognitive Psychology, 14,* 215–235. doi:10.1002/(SICI)1099-0720(200005/06)14:3<215::AID-ACP640>3.0.CO;2-1

Delaney, P. F., Verkoeijen, P. P. J. L., & Spirgel, A. (2010). Spacing and testing effects: A deeply critical, lengthy, and at times discursive review of the literature. *Psychology of Learning and Motivation, 53,* 63–147. doi:10.1016/S0079-7421(10)53003-2

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

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

Erlebacher, A. (1977). Design and analysis of experiments contrasting the within- and between-subjects manipulation of the independent variable. *Psychological Bulletin, 84,* 212–219. doi:10.1037/0033-2909.84.2.212

*Fadler, C. L., Bugg, J. M., & McDaniel, M. A. (2012). *The testing effect with authentic educational materials: A cautionary note*. Unpublished manuscript.

Ferguson, C. J., & Brannick, M. T. (2012). Publication bias in psychological science: Prevalence, methods for identifying and controlling, and implications for the use in meta-analyses. *Psychological Methods, 17,* 120–128. doi:10.1037/a0024445

*Finley, J. R., Benjamin, A. S., Hays, M. J., Bjork, R. A., & Kornell, N. (2011). Benefits of accumulating versus diminishing cues in recall. *Journal of Memory and Language, 64,* 289–298. doi:10.1016/j.jml.2011.01.006

*Fritz, C. O., Morris, P. E., Acton, M., Voelkel, A. R., & Etkind, R. (2007). Comparing and combining retrieval practice and the keyword mnemonic for foreign vocabulary learning. *Applied Cognitive Psychology, 21,* 499–526. doi:10.1002/acp.1287

Gates, A. I. (1917). Recitation as a factor in memorizing. *Archives of Psychology, 6*(40).

Gingerich, K. J., Bugg, J. M., Doe, S. R., Rowland, C. A., Richards, T. L., Tompkins, S. A., & McDaniel, M. A. (in press). Active processing during write-to-learn assignments produces learning and retention benefits in a large introductory psychology course. *Teaching of Psychology*.

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

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

Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics, 6,* 107–128. doi:10.2307/1164588

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

Higgins, J. P. T., Thompson, S. G., Deeks, J. J., & Altman, D. G. (2003). Measuring inconsistency in meta-analyses. *British Medical Journal, 327,* 557–560. doi:10.1136/bmj.327.7414.557

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

Hunt, R. R., & McDaniel, M. A. (1993). The enigma of organization and distinctiveness. *Journal of Memory and Language, 32,* 421–445. doi:10.1006/jmla.1993.1023

Hunter, J. E., & Schmidt, F. L. (2000). Fixed effects vs. random effects meta-analysis models: Implications for cumulative research knowledge. *International Journal of Selection and Assessment, 8,* 275–292. doi:10.1111/1468-2389.00156

Ioannidis, J. P. A., & Trikalinos, T. A. (2007). The appropriateness of asymmetry tests for publication bias in meta-analyses: A large survey. *Canadian Medical Association Journal, 176,* 1091–1096. doi:10.1503/cmaj.060410

Jacoby, L. L. (1978). On interpreting the effects of repetition: Solving a problem versus remembering a solution. *Journal of Verbal Learning and Verbal Behavior, 17,* 649–667. doi:10.1016/S0022-5371(78)90393-6

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

Jang, Y., Wixted, J. T., Pecher, D., Zeelenberg, R., & Huber, D. E. (2012). Decomposing the interaction between retention interval and study/test practice. *Quarterly Journal of Experimental Psychology, 65,* 962–975. doi:10.1080/17470218.2011.638079

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

*Kang, S. H. (2010). Enhancing visuospatial learning: The benefit of retrieval practice. *Memory & Cognition, 38,* 1009–1017. doi:10.3758/MC.38.8.1009

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

Karpicke, J. D., & Bauernschmidt, A. (2011). Spaced retrieval: Absolute spacing enhances learning regardless of relative spacing. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 37,* 1250–1257. doi:10.1037/a0023436

*Karpicke, J. D., & Blunt, J. R. (2011). Retrieval practice produces more learning than elaborative studying with concept mapping. *Science, 331,* 772–775. doi:10.1126/science.1199327

Karpicke, J. D., & Grimaldi, P. J. (2012). Retrieval-based learning: A perspective for enhancing meaningful learning. *Educational Psychology Review, 24,* 401–418. doi:10.1007/s10648-012-9202-2

Karpicke, J. D., & Roediger, H. L., III. (2007). Expanding retrieval practice promotes short-term retention, but equally spaced retrieval enhances long-term retention. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33,* 704–719. doi:10.1037/0278-7393.33.4.704

Karpicke, J. D., & Roediger, H. L., III. (2008). The critical importance of retrieval for learning. *Science, 319,* 966–968. doi:10.1126/science.1152408

*Karpicke, J. D., & Zaromb, F. M. (2010). Retrieval mode distinguished the testing effect from the generation effect. *Journal of Memory and Language, 62,* 227–239. doi:10.1016/j.jml.2009.11.010

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

*Kornell, N., & Son, L. K. (2009). Learners' choices and beliefs about self-testing. *Memory, 17,* 493–501. doi:10.1080/09658210902832915

*Kuo, T., & Hirshman, E. (1996). Investigations of the testing effect. *American Journal of Psychology, 109,* 451–464. doi:10.2307/1423016

Kuo, T., & Hirshman, E. (1997). The role of distinctive perceptual information in memory: Studies of the testing effect. *Journal of Memory and Language, 36,* 188–201. doi:10.1006/jmla.1996.2486

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

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

Little, J. L., Storm, B. C., & Bjork, E. L. (2011). The costs and benefits of testing text materials. *Memory, 19,* 346–359. doi:10.1080/09658211.2011.569725

*Littrell, M. K. (2008). *The effect of testing on false memory: Tests of the multiple-cue and distinctiveness explanations of the testing effect* (Unpublished master's thesis). Colorado State University, Fort Collins.

*Littrell, M. K. (2011). *The influence of testing on memory, monitoring, and control* (Unpublished doctoral dissertation). Colorado State University, Fort Collins.

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

*McConnell, M. D., & Hunt, R. R. (2007). Can false memories be corrected by feedback in the DRM paradigm? *Memory & Cognition, 35,* 999–1006. doi:10.3758/BF03193472

McDaniel, M. A., Agarwal, P. K., Huelser, B. J., McDermott, K. B., & Roediger, H. L., III. (2011). Testing-enhanced learning in a middle school science classroom: The effects of quiz frequency and placement. *Journal of Educational Psychology, 103,* 399–414. doi:10.1037/a0021782

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

McDaniel, M. A., & Bugg, J. M. (2008). Instability in memory phenomena: A common puzzle and a unifying explanation. *Psychonomic Bulletin & Review, 15,* 237–255. doi:10.3758/PBR.15.2.237

McDaniel, M. A., & Fisher, R. P. (1991). Tests and test feedback as learning sources. *Contemporary Educational Psychology, 16,* 192–201. doi:10.1016/0361-476X(91)90037-L

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

McDaniel, M. A., Roediger, H. L., III, & McDermott, K. B. (2007). Generalizing test-enhanced learning from the laboratory to the classroom. *Psychonomic Bulletin & Review, 14,* 200–206. doi:10.3758/BF03194052

McDermott, K. B. (2006). Paradoxical effects of testing: Repeated retrieval attempts enhance the likelihood of later accurate and false recall. *Memory & Cognition, 34,* 261–267. doi:10.3758/BF03193404

Metcalfe, J., Kornell, N., & Finn, B. (2009). Delayed versus immediate feedback in children's and adults' vocabulary learning. *Memory & Cognition, 37,* 1077–1087. doi:10.3758/MC.37.8.1077

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

Modigliani, V. (1976). Effects on a later recall by delaying initial recall. *Journal of Experimental Psychology: Human Learning and Memory, 2,* 609–622. doi:10.1037/0278-7393.2.5.609

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. doi:10.1016/S0022-5371(77)80016-9

*Morris, P. E., Fritz, C. O., Jackson, L., Nichol, E., & Roberts, E. (2005). Strategies for learning proper names: Expanding retrieval practice, meaning and imagery. *Applied Cognitive Psychology, 19,* 779–798. doi:10.1002/acp.1115

Nairne, J. S., Riegler, G. J., & Serra, M. (1991). Dissociative effects of generation on item and order information. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 17,* 702–709. doi:10.1037/0278-7393.17.4.702

*Neuschatz, J. S., Preston, E. L., Toglia, M. P., & Neuschatz, J. S. (2005). Comparison of the efficacy of two name-learning techniques: Expanding rehearsal and name-face imagery. *American Journal of Psychology, 118,* 79–101.

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. doi:10.1080/09658211.2011.648198

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

Odegard, T. N., & Koen, J. D. (2007). "None of the above" as a correct and incorrect alternative on a multiple-choice test: Implications for the testing effect. *Memory, 15,* 873–885. doi:10.1080/09658210701746621

*Peterson, D. J. (2011). *The testing effect and the item specific vs. relational account* (Unpublished doctoral dissertation). University of North Carolina at Chapel Hill.

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

Potts, R., & Shanks, D. R. (2012). Can testing immunize memories against interference? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 38,* 1780–1785. doi:10.1037/a0028218

*Putnam, A. L., & Roediger, H. L., III. (2013). Does response mode affect the amount recalled of the magnitude of the testing effect? *Memory & Cognition, 41,* 36–48. doi:10.3758/s13421-012-0245-x

Pyc, M. A., & Rawson, K. A. (2007). Examining the efficiency of schedules of distributed retrieval practice. *Memory & Cognition, 35,* 1917–1927. doi:10.3758/BF03192925

Pyc, M. A., & Rawson, K. A. (2009). Testing the retrieval effort hypothesis: Does greater difficulty correctly recalling information lead to higher levels of memory? *Journal of Memory and Language, 60,* 437–447. doi:10.1016/j.jml.2009.01.004

*Pyc, M. A., & Rawson, K. A. (2010). Why testing improves memory: Mediator effectiveness hypothesis. *Science, 330,* 335. doi:10.1126/science.1191465

*Pyc, M. A., & Rawson, K. A. (2011). Costs and benefits of dropout schedules of test-restudy practice: Implications for student learning. *Applied Cognitive Psychology, 25,* 87–95. doi:10.1002/acp.1646

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

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

Reyna, V. F., & Brainerd, C. J. (1995). Fuzzy-trace theory: An interim synthesis. *Learning and Individual Differences, 7,* 1–75. doi:10.1016/1041-6080(95)90031-4

Roediger, H. L., III, Agarwal, P. K., Kang, S. H. K., & Marsh, E. J. (2010). Benefits of testing memory: Best practices and boundary conditions. In G. M. Davies & D. B. Wright (Eds.), *New frontiers in applied memory* (pp. 13–49). Brighton, England: Psychology Press.

Roediger, H. L., III, Agarwal, P. K., McDaniel, M. A., & McDermott, K. B. (2011). Test-enhanced learning in the classroom: Long-term improvements from quizzing. *Journal of Experimental Psychology: Applied, 17,* 382–395. doi:10.1037/a0026252

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

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

*Roediger, H. L., III, & Karpicke, J. D. (2006b). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science, 17,* 249–255. doi:10.1111/j.1467-9280.2006.01693.x

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

Roediger, H. L., III, & McDermott, K. B. (1995). Creating false memories: Remembering words not presented in lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21,* 803–814. doi:10.1037/0278-7393.21.4.803

Roediger, H. L., III, Putnam, A. L., & Smith, M. A. (2011). Ten benefits of testing and their applications to educational practice. In J. Mestre & B. Ross (Eds.), *Psychology of learning and motivation: Cognition in education* (pp. 1–36). Oxford, England: Elsevier. doi:10.1016/B978-0-12-387691-1.00001-6

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

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

Rothstein, H. R., & Bushman, B. J. (2012). Publication bias in psychological science: Comment on Ferguson and Brannick (2012). *Psychological Methods, 17,* 129–136. doi:10.1037/a0027128

*Rowland, C. A. (2011). *Testing effects in context memory* (Unpublished master's thesis). Colorado State University, Fort Collins.

Rowland, C. A., & DeLosh, E. L. (2014a). Benefits of testing for nontested information: Retrieval-induced facilitation of episodically bound material. *Psychonomic Bulletin & Review*. doi:10.3758/s13423-014-0625-2

*Rowland, C. A., & DeLosh, E. L. (2014b). Mnemonic benefits of retrieval practice at short retention intervals. *Memory*. Advance online publication. doi:10.1080/09658211.2014.889710

*Rowland, C. A., Littrell-Baez, M. K., Sensenig, A. E., & DeLosh, E. L. (2014). Testing effects in mixed- versus pure-list designs. *Memory & Cognition*. Advance online publication. doi:10.3758/s13421-014-0404-3

Runquist, W. N. (1983). Some effects of remembering on forgetting. *Memory & Cognition, 11,* 641–650. doi:10.3758/BF03198289

Runquist, W. N. (1986). The effect of testing on the forgetting of related and unrelated associates. *Canadian Journal of Psychology, 40,* 65–76. doi:10.1037/h0080086

Schmidt, F. L., Oh, I., & Hayes, T. L. (2009). Fixed- versus random-effects models in meta-analysis: Model properties and an empirical comparison of differences in results. *British Journal of Mathematical and Statistical Psychology, 62,* 97–128. doi:10.1348/000711007X255327

*Sensenig, A. E. (2010). *Multiple choice testing and the retrieval hypothesis of the testing effect* (Unpublished doctoral dissertation). Colorado State University, Fort Collins.

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

Serra, M., & Nairne, J. S. (1993). Design controversies and the generation effect: Support for an item-order hypothesis. *Memory & Cognition, 21,* 34–40. doi:10.3758/BF03211162

Slamecka, N. J., & Katsaiti, L. T. (1987). The generation effect as an artifact of selective displaced rehearsal. *Journal of Memory and Language, 26,* 589–607. doi:10.1016/0749-596X(87)90104-5

*Smith, D. L. (2008). *The testing effect and the components of recognition memory: What effects do test type and performance at intervening test have on final recognition tests?* (Unpublished doctoral dissertation). Auburn University, Auburn, AL.

Smith, T. A., & Kimball, D. R. (2010). Learning from feedback: Spacing and the delay-retention effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36,* 80–95. doi:10.1037/a0017407

Spitzer, H. F. (1939). Studies in retention. *Journal of Educational Psychology, 30,* 641–656. doi:10.1037/h0063404

Storm, B. C., & Levy, B. J. (2012). A progress report on the inhibitory account of retrieval-induced forgetting. *Memory & Cognition, 40,* 827–843. doi:10.3758/s13421-012-0211-7

*Sumowski, J. F., Chiaravalloti, N., & DeLuca, J. (2010). Retrieval practice improves memory in multiple sclerosis: Clinical application of the testing effect. *Neuropsychology, 24,* 267–272. doi:10.1037/a0017533

Sutton, A. J. (2009). Publication bias. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis* (2nd ed., pp. 435–452). New York, NY: Russell Sage Foundation.

Szpunar, K. K., McDermott, K. B., & Roediger, H. L., III. (2008). Testing during study insulates against the build-up of proactive interference. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34,* 1392–1399. doi:10.1037/a0013082

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

*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. doi:10.1037/a0028886

*Thompson, C. P., Wenger, S. K., & Bartling, C. A. (1978). How recall facilitates subsequent recall: A reappraisal. *Journal of Experimental Psychology: Human Learning and Memory, 4,* 210–221. doi:10.1037/0278-7393.4.3.210

*Toppino, T. C., & Cohen, M. S. (2009). The testing effect and the retention interval: Questions and answers. *Experimental Psychology, 56,* 252–257. doi:10.1027/1618-3169.56.4.252

Vaughn, K. E., & Rawson, K. A. (2011). Diagnosing criterion level effects on memory: What aspects of memory are enhanced by repeated retrieval? *Psychological Science, 22,* 1127–1131. doi:10.1177/0956797611417724

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

*Verkoeijen, P. P. J. L., & Delaney, P. F. (2012). *Encoding strategy and the testing effect in free recall*. Unpublished manuscript.

*Verkoeijen, P. P. J. L., Delaney, P. F., Bouwmeester, S., Coppens, L. C., & Spirgel, A. (2012). *No testing effect in categorized lists: Using a Bayesian approach to support the null hypothesis*. Unpublished manuscript.

*Wartenweiler, D. (2011). Testing effect for visual-symbolic material: Enhancing the learning of Filipino children of low socio-economic status in the public school system. *International Journal of Research and Review, 6,* 74–93.

Weinstein, Y., McDermott, K. B., & Szpunar, K. K. (2011). Testing protects against proactive interference in face–name learning. *Psychonomic Bulletin & Review, 18,* 518–523. doi:10.3758/s13423-011-0085-x

*Wheeler, M. A., Ewers, M., & Buonanno, J. F. (2003). Different rates of forgetting following study versus test trials. *Memory, 11,* 571–580. doi:10.1080/09658210244000414

Whitten, W. B., II, & Bjork, R. A. (1977). Learning from tests: Effects of spacing. *Journal of Verbal Learning and Verbal Behavior, 16,* 465–478. doi:10.1016/S0022-5371(77)80040-6

Wood, W., & Eagly, A. H. (2009). Advantages of certainty and uncertainty. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis* (2nd ed., pp. 455–472). New York, NY: Russell Sage Foundation.

*Zaromb, F. M., & Roediger, H. L., III. (2010). The testing effect in free recall is associated with enhanced organizational processes. *Memory & Cognition, 38,* 995–1008. doi:10.3758/MC.38.8.995

(*Appendices follow*)

Appendix A

Studies Included in the Meta-Analysis Presented With Selected Moderator Classifications

Study	Stimulus type	Initial test type	Initial test cue–target relationship	Feedback	Retention interval	Final test type
Bishara & Jacoby (2008)						
Experiment 1	PA	CR	SR	Yes	2.8	CR
Experiment 1	PA	CR	SR	Yes	2.8	CR
Brewer & Unsworth (2012)						
Experiment 1	PA	CR	SR	Yes	1,440	CR
Butler (2010)						
Experiment 1a	Prose	CR	SR	Yes	10,080	CR
Carpenter (2009)						
Experiment 1	PA	CR	SR	No	5	FR
Experiment 2	PA	CR	SR	No	5	FR
Experiment 2	PA	CR	SR	No	5	FR
Carpenter (2011)						
Experiment 1	PA	CR	SR	No	5	Rec
Experiment 2	PA	CR	SR	No	30	CR
Carpenter & DeLosh (2005)						
Experiment 1a	PA	CR	NS	No	5	CR
Experiment 1b	PA	CR	NS	No	5	CR
Experiment 2	PA	CR	NS	No	5	CR
Experiment 3	PA	CR	NS	No	5	CR
Carpenter & DeLosh (2006)						
Experiment 1	SW	Rec	Same	No	5	Rec
Experiment 1	SW	FR	None	No	5	CR
Experiment 1	SW	Rec	Same	No	5	FR
Experiment 2	SW	CR	NS	No	5	FR
Carpenter & Pashler (2007)						
Experiment 1	O	CR	NS	Yes	30	FR
Carpenter et al. (2006)						
Experiment 1	PA	CR	SR	Yes	1,980	CR
Experiment 1	PA	CR	SR	Yes	1,980	FR
Experiment 2	PA	CR	SR	Yes	1,980	CR
Experiment 2	PA	CR	SR	Yes	1,980	FR
Carpenter et al. (2008)						
Experiment 1	O	CR	SR	Yes	20,160	CR
Experiment 2	O	CR	SR	Yes	5	CR
Experiment 3	PA	CR	NS	Yes	5	CR
Carrier & Pashler (1992)						
Experiment 4	PA	CR	NS	Yes	2	CR
Chan et al. (2006)						
Experiment 1	Prose	CR	SR	No	1,440	CR
Congleton & Rajaram (2012)						
Experiment 1	SW	FR	None	No	7	FR
Experiment 1	SW	FR	None	No	10,080	FR
Coppens et al. (2011)						
Experiment 1	PA	CR	NS	No	5	CR
Experiment 1	PA	CR	NS	No	10,080	CR
Cull (2000)						
Experiment 1	PA	CR	SU	Yes	1	CR
Fadler et al. (2012)						
Experiment 1	Prose	CR	SR	Yes	2,880	Rec
Finley et al. (2011)						
Experiment 1	PA	CR	NS	No	10	CR
Experiment 2	PA	CR	NS	Yes	10	CR
Fritz et al. (2007)						
Experiment 1	PA	CR	NS	Yes	3	FR
Halamish & Bjork (2011)						
Experiment 1	PA	CR	SR	No	1.5	CR
Experiment 1	PA	CR	SR	No	1.5	CR

(*Appendices continue*)

Appendix A (*continued*)

Study	Stimulus type	Initial test type	Initial test cue–target relationship	Feedback	Retention interval	Final test type
Experiment 1	PA	CR	SR	No	1.5	FR
Experiment 2	PA	CR	SR	No	1.5	CR
Experiment 2	PA	CR	SR	No	1.5	FR
Hinze & Wiley (2011)						
Experiment 1	Prose	CR	SR	No	2,880	CR
Experiment 2	Prose	CR	SR	No	10,080	CR
Jacoby et al. (2010)						
Experiment 1	PA	CR	NS	Yes	0	Rec
Experiment 2	PA	CR	NS	Yes	0	Rec
Experiment 3	PA	CR	NS	No	0	Rec
Experiment 3	PA	CR	NS	No	1,440	Rec
Johnson & Mayer (2009)						
Experiment 1	O	FR	None	No	3	FR
Experiment 1	O	FR	None	No	10,080	FR
Kang (2010)						
Experiment 1	PA	CR	NS	Yes	10	CR
Experiment 2	PA	CR	NS	Yes	1,440	CR
Experiment 3	PA	CR	NS	Yes	10	CR
Kang et al. (2007)						
Experiment 1	Prose	Rec	Same	No	4,320	Rec
Experiment 2	Prose	Rec	Same	Yes	4,320	Rec
Karpicke & Blunt (2011)						
Experiment 1	Prose	FR	None	Yes	10,080	CR
Karpicke & Zaromb (2010)						
Experiment 1	SW	CR	SR	No	5	FR
Experiment 2	SW	CR	SR	No	5	FR
Experiment 3	SW	CR	SR	No	5	Rec
Experiment 4	SW	CR	SR	No	5	FR
Kornell et al. (2011)						
Experiment 1	PA	CR	SR	No	2	CR
Experiment 2	PA	CR	SR	Yes	2	CR
Kornell & Son (2009)						
Experiment 1	PA	CR	SR	No	5	CR
Experiment 1	PA	CR	SR	Yes	5	CR
Kuo & Hirshman (1996)						
Experiment 1	SW	FR	None	No	5	FR
Experiment 1	SW	FR	None	No	10	FR
Laporte & Voss (1975)						
Experiment 1	Prose	CR	SR	Yes	10,080	CR
Littrell (2008)						
Experiment 1	SW	FR	None	No	5	FR
Experiment 2	PA	CR	SR	No	10	Rec
Experiment 3	PA	CR	SR	No	10	CR
Littrell (2011)						
Experiment 2	PA	CR	SU	No	5	CR
McConnell & Hunt (2007)						
Experiment 1	SW	FR	None	Yes	2,880	FR
Experiment 2	SW	FR	None	Yes	2,880	FR
Meyer & Logan (2013)						
Experiment 1	Prose	Rec	Same	No	5	CR
Experiment 1	Prose	Rec	Same	No	5	CR
Experiment 1	Prose	Rec	Same	No	5	CR
Experiment 1	Prose	Rec	Same	No	2,880	CR
Experiment 1	Prose	Rec	Same	No	2,880	CR
Experiment 1	Prose	Rec	Same	No	2,880	CR
P. E. Morris et al. (2005)						
Experiment 1	PA	CR	NS	No	5	CR
Experiment 1	PA	CR	NS	No	5	CR
Experiment 2	PA	CR	NS	Yes	5	CR
Experiment 2	PA	CR	NS	Yes	5	CR

(*Appendices continue*)

Appendix A (*continued*)

Study	Stimulus type	Initial test type	Initial test cue–target relationship	Feedback	Retention interval	Final test type
Neuschatz et al. (2005)						
Experiment 3	PA	CR	NS	No	15	CR
Nungester & Duchastel (1982)						
Experiment 1	Prose	CR	SR	No	20,160	CR
Peterson (2011)						
Experiment 1	PA	CR	SU	Yes	5	FR
Peterson & Mulligan (2013)						
Experiment 1	PA	CR	SU	Yes	0	FR
Experiment 2	PA	CR	SU	Yes	0	CR
Experiment 3	PA	CR	SU	Yes	0	FR
Putnam & Roediger (2013)						
Experiment 1	PA	CR	SR	No	2,880	CR
Experiment 2	PA	CR	SR	Yes	2,880	CR
Experiment 3	PA	CR	SR	Yes	2,880	CR
Pyc & Rawson (2010)						
Experiment 1	PA	CR	SU	Yes	10,080	CR
Experiment 1	PA	CR	SU	Yes	10,080	CR
Experiment 1	PA	CR	SU	Yes	10,080	CR
Pyc & Rawson (2011)						
Experiment 1a	PA	CR	SU	Yes	2,880	CR
Experiment 1b	PA	CR	SU	Yes	2,880	CR
Experiment 2	PA	CR	SU	Yes	2,880	CR
Roediger & Karpicke (2006b)						
Experiment 1	Prose	FR	None	No	5	FR
Experiment 1	Prose	FR	None	No	2,880	FR
Experiment 1	Prose	FR	None	No	10,080	FR
Experiment 2	Prose	FR	None	No	10,080	FR
Experiment 2	Prose	FR	None	No	10,080	FR
Rohrer et al. (2010)						
Experiment 1	O	Rec	Same	Yes	1,440	Rec
Experiment 1	O	Rec	Same	Yes	1,440	Rec
Rowland (2011)						
Experiment 1	SW	CR	NS	No	4	Rec
Experiment 2	PA	CR	SU	No	4	CR
Rowland & DeLosh (2014b)						
Experiment 1	SW	CR	NS	No	4	FR
Experiment 2	SW	CR	NS	No	8	FR
Experiment 3	SW	CR	NS	No	0.5	FR
Experiment 3	SW	CR	NS	No	1.5	FR
Experiment 3	SW	CR	NS	No	4	FR
Experiment 4	SW	CR	NS	No	0.5	FR
Experiment 4	SW	CR	NS	No	1.5	FR
Rowland et al. (2014)						
Experiment 1	SW	CR	NS	No	5	FR
Experiment 1	SW	CR	NS	No	5	FR
Experiment 1	SW	CR	NS	No	5	FR
Experiment 1	SW	CR	NS	No	5	FR
Experiment 2	SW	CR	NS	No	5	FR
Experiment 3	SW	CR	NS	No	4	FR
Experiment 3	SW	CR	NS	No	4	FR
Sensenig (2010)						
Experiment 1	Prose	Rec	Same	No	5	CR
Experiment 2a	Prose	Rec	Same	No	5	CR
Sensenig et al. (2011)						
Experiment 1	SW	CR	NS	No	5	CR
Experiment 1	SW	CR	NS	No	5	CR
Experiment 2	SW	FR	None	No	5	CR
Experiment 2	SW	FR	None	No	5	CR
Experiment 3	SW	CR	NS	No	5	CR
Experiment 3	SW	CR	NS	No	5	CR

(*Appendices continue*)

Appendix A (*continued*)

Study	Stimulus type	Initial test type	Initial test cue–target relationship	Feedback	Retention interval	Final test type
D. L. Smith (2008)						
Experiment 1	SW	Rec	Same	No	3	Rec
Experiment 1	SW	Rec	Same	No	3	Rec
Experiment 2	SW	Rec	Same	No	3	Rec
Experiment 2	SW	Rec	Same	No	3	Rec
Sumowski et al. (2010)						
Experiment 1	PA	CR	SR	Yes	45	CR
Thomas & McDaniel (2013)						
Experiment 1	PA	CR	SR	Yes	2,880	CR
Experiment 2	PA	CR	SR	No	2,880	CR
Experiment 2	PA	CR	SR	Yes	2,880	CR
Thompson et al. (1978)						
Experiment 3	SW	FR	None	No	2,880	FR
Toppino & Cohen (2009)						
Experiment 1	PA	CR	NS	No	2	CR
Experiment 1	PA	CR	NS	No	2,880	CR
Experiment 2	PA	CR	NS	No	5	CR
Experiment 2	PA	CR	NS	No	2,880	CR
Verkoeijen, Bouwmeester, & Camp (2012)						
Experiment 1	SW	FR	None	No	2	Rec
Experiment 1	SW	FR	None	No	2	Rec
Verkoeijen & Delaney (2012)						
Experiment 1	SW	FR	None	No	5	FR
Experiment 1	SW	FR	None	No	10,080	FR
Experiment 1	SW	FR	None	No	5	FR
Experiment 1	SW	FR	None	No	10,080	FR
Verkoeijen, Delaney, et al. (2012)						
Experiment 1	SW	FR	None	No	10,080	FR
Experiment 2	SW	FR	None	No	5	FR
Experiment 2	SW	FR	None	No	5	FR
Experiment 2	SW	FR	None	No	10,080	FR
Experiment 2	SW	FR	None	No	10,080	FR
Wartenweiler (2011)						
Experiment 1	PA	Rec	Same	Yes	60	Rec
Wheeler et al. (2003)						
Experiment 1	SW	FR	None	No	5	FR
Experiment 1	SW	FR	None	No	2,880	FR
Experiment 2	SW	FR	None	No	5	FR
Experiment 2	SW	FR	None	No	10,080	FR
Zaromb & Roediger (2010)						
Experiment 1	SW	FR	None	Yes	2,880	FR
Experiment 1	SW	FR	None	No	1,440	FR

Note. Retention interval durations indicated in minutes. SW = single words; PA = paired associates; O = other materials; CR = cued recall; FR = free recall; Rec = recognition; NS = nonsemantic; SR = semantic related; SU = semantic unrelated.

(*Appendices continue*)

Appendix B

Supplementary Data Set Analyses

Categorical moderator analyses are reported for a restricted data set of studies utilizing long retention intervals (at least 1 day) in Table B1. Note that the number of effect sizes is lower than in the full and high-exposure data sets; thus, caution is encouraged when interpreting the analyses given the relatively lower power and, in some cases, small cell sizes.

Table B1
Long Retention Interval Studies Data Set Categorical Moderator Analyses

Moderator	*g*	95% CI *LL*	95% CI *UL*	Q_B	*k*
Publication status				5.62*	
Published	0.72	0.61	0.86		50
Unpublished	0.20	−0.22	0.63		6
Sample source				1.39	
College	0.72	0.56	0.87		45
Other	0.58	0.41	0.74		11
Design				13.72**	
Between	0.95	0.76	1.14		26
Within	0.50	0.36	0.65		30
Stimulus type				6.24	
Prose	0.76	0.53	1.00		17
Paired associates	0.70	0.53	0.86		22
Single words	0.66	0.24	1.07		13
Other	0.40	0.19	0.61		4
Stimulus interrelation				2.71	
Prose	0.76	0.53	1.00		17
Categorical	0.84	0.30	1.38		8
No relation	0.62	0.47	0.78		28
Other	0.52	0.30	0.74		3
List blocking				3.72	
Mixed	0.44	0.20	0.69		5
Blocked	0.70	0.52	0.89		33
Other	0.74	0.52	0.96		18
Initial test type				1.41	
Cued recall	0.69	0.54	0.84		30
Free recall	0.74	0.45	1.04		19
Recognition	0.51	0.22	0.81		7
Initial test cue–target relationship				9.01	
Same (recognition)	0.51	0.22	0.81		7
Nonsemantic	0.60	0.32	0.88		5
Semantic unrelated	1.02	0.77	1.28		6
Semantic related	0.62	0.44	0.81		19
None (free recall)	0.74	0.45	1.04		19
Final test type				4.39	
Cued recall	0.75	0.59	0.90		30
Free recall	0.68	0.41	0.95		20
Recognition	0.40	0.11	0.69		6
Initial–final test match				0.05	
Different	0.66	0.40	0.91		5
Same	0.69	0.55	0.83		48
Feedback				6.16*	
No	0.53	0.36	0.70		29
Yes	0.85	0.67	1.03		27
Retrievability and reexposure				13.50	
No feedback ≤50%	0.26	0.01	0.52		7
No feedback 51%–75%	0.61	0.17	1.15		9
No feedback >75%	0.57	0.28	0.86		8
Feedback	0.85	0.67	1.03		27
Unknown and no feedback	0.68	0.44	0.92		5

Note. *g* = mean weighted effect size; CI = confidence interval; *LL* = lower limit; *UL* = upper limit; *k* = number of effect sizes.
* Q_B test for heterogeneity between levels of a moderator was significant at $p < .05$. ** Q_B test for heterogeneity between levels of a moderator was significant at $p < .01$.

(*Appendices continue*)

Appendix C

Descriptive Contingency Tables for the High-Exposure Data Set

Contingency tables for select variables of interest are presented in Table C1 for the effect sizes in the high-exposure data set. Each table indicates the number of effect sizes belonging to a specific group in one moderator variable, as distributed across the groups of a second moderator variable.

Table C1
Contingency Tables for the High-Exposure Data Set

Moderator	Level	Rec	CR	FR	Same	Different	Total
Initial–Final Test Match × Initial Test Type	Same	10	36	7			39
	Different	7	29	3			53
	Total	17	65	10			92
Initial–Final Test Match × Final Test Type	Same	10	36	7			53
	Different	7	7	25			39
	Total	17	43	32			92
Feedback × Initial Test Type	Yes	4	44	4			52
	No	13	21	6			40
	Total	17	65	10			92
Feedback × Final Test Type	Yes	7	33	12			52
	No	10	10	20			40
	Total	17	43	32			92
Feedback × Initial–Final Test Match	Yes				39	13	52
	No				14	26	40
	Total				53	39	92

Note. Rec = recognition; CR = cued recall; FR = free recall.

Received November 12, 2012
Revision received April 5, 2014
Accepted May 21, 2014 ■