Psychon Bull Rev
DOI 10.3758/s13423-017-1298-4

THEORETICAL REVIEW

A dual memory theory of the testing effect

Timothy C. Rickard[1,2] · Steven C. Pan[1,2]

Abstract A new theoretical framework for the *testing effect*—the finding that retrieval practice is usually more effective for learning than are other strategies—is proposed, the empirically supported tenet of which is that separate memories form as a consequence of study and test events. A simplest case quantitative model is derived from that framework for the case of cued recall. With no free parameters, that model predicts both proportion correct in the test condition and the magnitude of the testing effect across 10 experiments conducted in our laboratory, experiments that varied with respect to material type, retention interval, and performance in the restudy condition. The model also provides the first quantitative accounts of (a) the testing effect as a function of performance in the restudy condition, (b) the upper bound magnitude of the testing effect, (c) the effect of correct answer feedback, (d) the testing effect as a function of retention interval for the cases of feedback and no feedback, and (e) the effect of prior learning method on subsequent learning through testing. Candidate accounts of several other core phenomena in the literature, including test-potentiated learning, recognition versus cued recall training effects, cued versus free recall final test effects, and other select transfer effects, are also proposed. Future prospects and relations to other theories are discussed.

Keywords Retrieval practice · Testing effect · Test-enhanced learning · Memory · Quantitative model

✉ Timothy C. Rickard
trickard@ucsd.edu

[1] University of California, San Diego, La Jolla, CA, USA

[2] Department of Psychology, University of California, San Diego, La Jolla, CA 92093-0109, USA

Introduction

Retrieval from long-term memory improves subsequent test performance more than does either no reexposure or an equivalent period of time allocated to other learning strategies (Bjork, 1975; Carrier & Pashler, 1992; Gates, 1917; Glover, 1989; Roediger & Butler, 2011). That memorial benefit is known as the *retrieval practice effect, test-enhanced learning,* and, in this article, the *testing effect* (*TE*).

Typically, the *TE* is explored using a three-phase ex>perimental design, involving (a) an *initial study phase* for items such as paired associates or biology facts, (b) a *training phase* (also referred to in the literature as the practice or reexposure phase) in which half of the items are *restudied* and half undergo an *initial test* (e.g., with one word of a paired associate presented as a retrieval cue for the other word), and (c) a *final test phase* in which all items are tested. The *TE*—usually measured quantitatively as final test proportion correct in the test condition minus that in the restudy condition—has been well-established in memory domains ranging from verbal to visuospatial (e.g., McDermott, Agarwal, D'Antonio, Roediger, & McDaniel, 2014; Rohrer, Taylor, & Sholar, 2010), with published demonstrations now numbering well over 200 across 150 papers and growing (Rawson & Dunlosky, 2011; Rowland, 2014). It is observed after both short (e.g., 1-minute) and extended (e.g., several month) retention intervals between the training and final test phases (e.g., Carpenter, Pashler, & Cepeda, 2009; Rowland & DeLosh, 2015), and the provision of correct answer feedback (henceforth, *feedback*) after each retrieval attempt increases the effect (e.g., Pashler, Cepeda, Wixted, & Rohrer, 2005). Given that preponderance of evidence, the *TE* now ranks among the most robust of psychological phenomena, and retrieval practice is considered to be among the most promising techniques for improving learning in educational contexts

Published online: 05 June 2017

(e.g., Dunlosky, Rawson, Marsh, Nathan, & Willingham, 2013; Roediger & Pyc, 2012).

Several theories have been proposed that are applicable to the testing effect, each specifying a mechanism that may prove to be important in understanding it. They include (1) the *desirable difficulties model* (Bjork, 1994), according to which more difficult retrieval yields more learning than does either restudy or less difficult retrieval; (2) the *distribution-based bifurcation model* (Kornell, Bjork, & Garcia, 2011; Halamish & Bjork, 2011), according to which testing without feedback results in bifurcation of the memory strength distribution by response accuracy, potentially explaining the atypical negative *TE*s that are often observed at short (e.g., 5 min) of retention intervals when there is no feedback; (3) the *elaborative retrieval hypothesis* (Carpenter, 2009), according to which retrieval with feedback creates more associative paths between the cue and the correct response than does restudy; (4) the *mediator effectiveness hypothesis* (Pyc & Rawson, 2010), according to which cue-response mediators are more likely to be established on test than on restudy trials; (5) the *episodic context theory* (Karpicke, Lehman, & Aue, 2014), which proposes that differences in frequency of retrieval and the degree of episodic context updating in the restudy versus test conditions explain the *TE*; (6) the *gist-trace processing* account (Bouwmeester & Verkoeijen, 2011), according to which testing strengthens memory primarily at the semantic level whereas restudy strengthens memory for surface features; and (7) the *attenuated error correction theory* (Mozer, Howe, & Pashler, 2004), according to which error correction processes in a feed-forward two-layer neural network model explain the *TE* for the case of testing with feedback.

The majority of those theories are expressed in conceptual terms, having not yet been implemented quantitatively, with Mozer et al.'s (2004) attenuated error correction theory constituting the primary exception. To date, none of those theories makes direct quantitative predictions about either proportion correct in the test condition or the *TE* magnitude. Here we propose a simple but powerful new theory, along with a corresponding quantitative model, that makes predictions for both test condition performance and the *TE* magnitude. We show that the model, which in its simplest case has no free parameters, can provide good-to-excellent quantitative accounts of several phenomena that are central to the *TE* literature, including but not limited to (a) the magnitude of the *TE* as a function of performance in the restudy condition, (b) the upper bound magnitude of the *TE*, (c) the effect of feedback, (d) the functional form of the *testing retention curve* (i.e., the *TE* magnitude as a function of the retention interval between training and final test phases) for the cases of both feedback and no feedback, and (e) the effect of training type on subsequent learning through testing with feedback.

A dual memory theoretical framework

The basic and unique claim of our theory is that the initial study phase encodes *study memory*, restudy strengthens study memory, and initial testing both strengthens study memory and encodes a new and separate *test memory.* Hence, final test performance in the restudy condition is supported solely by study memory, whereas final test performance in the test condition can be supported by study memory, test memory, or both.

Our assumption that a restudy trial reactivates and strengthens the originally encoded study memory is consistent with recent work showing that, for at least the case of pure repetition, *study phase retrieval* (Hintzman, 2010; Thios & D'Agostino, 1976) is a core phenomenon that should be included in any general theory of the spacing effect. Benjamin and Tullis (2010), for example, marshalled evidence through both meta-analyses and model development suggesting that the second presentation of an item will, with a probability associated with the delay between presentations, "remind" (i.e., reactivate) and "enhance" (i.e., strengthen) the earlier encoded study memory. Based on their meta-analysis, reminding appears to occur frequently at repetition lags of up to 80 trials or more, which exceeds the mean item repetition lag between initial study and restudy in the majority of cued recall *TE* literature. Although those results do not guarantee that restudy in *TE* experiments will always result in reminding and strengthening of prior study memory, for simplicity we assume that it always does in the model development below.

Now consider initial test trials. Our claim that an initial test trial with feedback reactivates and strengthens study memory is a natural extension of our assumption that restudy reactivates and strengthens study memory. Successful retrieval on the initial test trial must involve reactivation of the corresponding study episode (provided that no preexperimental associations that would support that retrieval exist), and that reactivation would be expected to strengthen that study memory, perhaps to roughly the same extent as does restudy. On incorrect initial test trials, the test cue plus the correct answer feedback reconstitute the full set of initial study stimulus elements, just as restudy does, and thus reactivation of study memory may occur during feedback even if the test cue alone was insufficient for that reactivation to occur. The consequent study memory strengthening may plausibly occur to roughly the same extent as it would have had that incorrectly answered item instead been in the restudy condition.

Critically in the dual memory model, the first test trial for an item also yields a new and separate *test memory.* Test memory has two components in the model: (1) *cue memory*, which is episodic encoding of the presented retrieval cue in the context of a task set to retrieve the response (as opposed to a presumed task set to memorize the full stimulus on study trials), and (2) an association between cue memory and the correct response. Cue memory is assumed to be encoded concurrently with cue presentation and independently of any subsequent answer retrieval attempt. When the correct response (e.g., the missing element of a paired associate) is then retrieved from study memory, or becomes available through feedback, an association between cue memory and that response can form, providing a second route to answer retrieval on later test trials.

Empirical evidence supporting separate study and test memory

Studies of process shifts during retrieval practice support our claim of separate study and test memory while also providing insight into the rate of shift from reliance on study memory to reliance on test memory. Because cue memory, once formed, is a better match to the presented cue and task set on subsequent test trials than is study memory, it is reasonable to expect that test memory ultimately will be the more optimal retrieval route, and that it will come to dominate the study memory route with sufficient practice.

One source of evidence for that shift comes from experiments in which a mnemonic mediator for retrieval is learned during initial study. For example, in some implementations of the keyword foreign vocabulary learning task (Atkinson & Raugh, 1975; Raugh & Atkinson, 1975), a foreign word (e.g., the French word *assiette*), an easy-to-recall, phonetically (or orthographically) related mediator word (keyword) from the native language (e.g., the English word *essay*), and the correct translation (e.g., *plate*) are all presented simultaneously for *study*. During that study phase, subjects are instructed to form an interactive image between the keyword and the foreign word (e.g., the act of writing an *essay* on a *plate*). In the experiments described below, subjects then practiced retrieving the English word when presented with only the foreign word (for related mediator tasks in the testing effect literature, see Pyc & Rawson, 2010, 2012).

Based on subject reports, the keyword mediator drops out of conscious use following retrieval practice (e.g., Crutcher & Ericsson, 2000), suggesting a shift from retrieval through study memory using the keyword mediator to a more direct retrieval process that bypasses that mediator (i.e., test memory). Kole and Healy (2013) provided direct evidence for that inference using a priming task. Their subjects performed a lexical decision task after some of the translation trials in which the word presented for lexical decision was either unrelated to or semantically related to the keyword. They observed a lexical priming advantage for the semantically related words at low translation practice levels, but not at moderate (five translation trials per item) or high (45 translation trials per item) practice levels. Translation retrieval practice thus appears to create a new retrieval route (i.e., through test memory) that bypasses keyword activation and thus appears to be functionally distinct from the memory that was formed by initial study.

The functional form of response time (RT) improvement with retrieval practice constitutes a second type of evidence. In the literature on practice effects, the power function best describes (and almost perfectly fits in most cases) RT improvement for averaged data across a wide array of tasks (Newell & Rosenbloom, 1981), including memory retrieval tasks that have been highly practiced prior to the experiment (e.g., single-digit arithmetic for college students, as in Rickard, 2007). For such tasks, it is unlikely that frequent and substantial strategy or process shifts occur during experimental practice. On the other hand, RT improvement on tasks that exhibit a clear strategy shift from reliance on a multistep algorithm to reliance on direct memory retrieval (e.g., novel multistep arithmetic tasks) does not follow a power function. Rather, RT improvement matches predictions of a mixture model (Bajic & Rickard, 2009; Delaney, Reder, Staszewski, & Ritter, 1998; Rickard, 1997) in which there is a shift for each item from one power function (describing RT improvement for the algorithm strategy) to a different power function (describing improvement for the retrieval strategy). Because the retrieval strategy power function has, as an empirical matter, faster RTs across the full range of practice than does the algorithm power function, the overall RT curve for such tasks is demonstrably not a power function but rather is an empirically distinct mixture of power functions.

Of interest here is the shape of the RT curve for the case of retrieval practice on a newly formed episodic memory, as in the case of an initial study trial followed by repeated test trials in test condition of the *TE* paradigm. If a single power function provides excellent fits to that RT curve from the first trial onward, then we can reasonably infer, based on findings summarized above, that no process shift occurs and that the memory that was encoded in the initial study phase continues to mediate performance. If, however, the curve is well fitted only by the mixture model, then a process shift is implied. In this case that result would suggest not a shift from a slow multistep algorithm to memory retrieval but rather a shift from relatively slow *study memory* access to reliance on a separate and ultimately more efficient retrieval route through *test memory.* Rickard and Bajic (2006) demonstrated that mixture model effect in each of three experiments in which there was initial study on a set of word triplets, followed by 20 blocks of retrieval practice in which the same two words of each triplet were presented once in each block as a retrieval cue for the third word. Hence, during retrieval practice (including the test condition of the *TE* paradigm), a shift from reliance on study memory to reliance on test memory appears to occur.

Both the Kole and Healy (2013) and Rickard and Bajic (2006) results suggest that test memory is the primary route to answer retrieval after about the first five to 10 test repetitions per item. It therefore appears that test memory develops quickly and that the two retrieval routes—through study memory and test memory—can jointly contribute to performance over at least the first several test trials per item. Because most studies in the *TE* literature involve one or only a few repetitions per item during training, the bulk of that literature appears to occupy the "sweet spot" within which retrieval contributions from both study and test memory would be expected for tested items on the final test.

Associative properties of study and test memory

The model currently makes no distinction between the associative properties of study memory and cue memory. They may both be instances of the same type of study memory. There is an important distinction in the model, however, between both of those instances of study memory and test memory. Study memory as conceived here has no necessary cue-response distinction. Rather, stimulus elements (including task set) can be bound together in the absence of that distinction. An example in the memory literature is the empirically supported schema model of Ross and Bower (1981), according to which study of a stimulus with multiple elements results in symmetric associative links between stimulus element nodes and a central "schema" node. Test memory, in contrast, involves formation of cue memory, and subsequently, formation of an association to a response. Under certain conditions, learning and strengthening of cue memory can occur independently of associative learning between that memory and the response. As we show in the General Discussion, those contrasting associative properties of study and test memory allow the model to uniquely account for important auxiliary phenomena in the *TE* literature.

A quantitative model based on the dual memory framework

The relatively simple dual memory framework outlined above can in principle explain core aspects of the *TE* in all contexts, a topic to which we will return in the General Discussion. The model described here applies to experiments in which both the initial and final tests involve cued recall, constituting roughly half of the experimental literature as catalogued by Rawson and Dunlosky (2011) and Rowland (2014). The model is also intended to apply most directly to the three phase experimental design outlined earlier, and to cases in which (a) episodic memory encoding during the initial study phase constitutes the only learning that can support answer retrieval on the initial test, (b) initial study involves a single trial per item of roughly the same latency as for both restudy and test trials in the training phase, and (c) correct answer feedback, if provided during training, is immediate after each test trial. Those conditions are, respectively, the most common in the cued recall *TE* literature.

In the following two sections, the model concepts are briefly elaborated and the equations are specified. Model development was guided at every decision point by the combined criteria of empirical support and theoretical precedent (where available) as well as model simplicity, a strategy that yielded a parameter-free model. We first evaluated the simplest case model that is consistent with the theoretical framework, and then considered possible elaboration as needed.

Conceptual description

Our primary interest here is not detailed modeling of study or test memory per se but rather development of a model that illustrates how the interaction of those two memories in the test condition may give rise to the testing effect. We thus adopt a strength-threshold account of both study memory and test memory. For current purposes, test memory strength corresponds to the combined influence of cue memory and the association to the response (distinctive roles of those two components will be considered later). For both study and test memory, a correct response is retrieved on a final test trial only if memory strength is above the response threshold. For strength-threshold model precedent in the *TE* literature, see Halamish and Bjork (2011) and Kornell et al. (2011); and in the memory literature more generally, see Wixted (2007).

Consider a hypothetical ideal subject with an infinite number of items, each randomly assigned to either the restudy or test condition. The initial study phase yields identical study memory strength probability distributions for items in those two conditions (see Fig. 1a–b; "after initial study"). For purposes of exposition and later simulation modeling, associative strengths are assumed to be gamma distributed. (However, none of the model predictions depend critically on that distribution assumption, and predictions for the case of testing with feedback are entirely independent of the shape of the memory strength distribution). The gamma distribution has a shape parameter, which is held constant at 2.0 for simulations in this article, and a scale parameter. In the figures and simulations, training effects are modeled by increasing the value of the scale parameter, which "stretches" the strength distribution to the right, increasing both its mean and standard deviation.

During the training phase, a restudy trial strengthens study memory for each item, resulting in the right-shifted strength distribution in Fig. 1a ("after training"). On testing with feedback trials—the type of test trial on which we focus first—two changes occur in memory: Study memory is strengthened, and test memory is encoded.

Given the conditions described earlier for which the model is developed, the first correct initial test trial must (excluding correct guessing) involve reactivation, and by our assumption strengthening, of study memory. On an incorrect trial, study

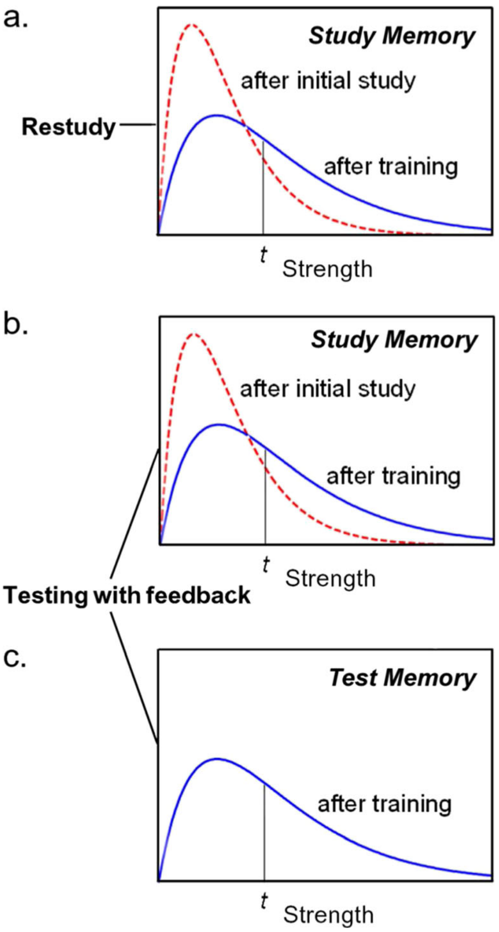

Fig. 1 Conceptual depiction of strength distributions under the dual memory model. **a** Gamma distributions of study memory strength in the *restudy* condition after initial study (*dashed line*) and after training (*solid line*). **b** and **c** Gamma distributions of study memory and test memory strength in the *testing with feedback* condition

memory can be reactivated and strengthened after feedback is presented. For simplicity, we assume here that study memory strengthening on test trials with feedback is not causally dependent on trial accuracy. We further assume that the amount of study memory strengthening on test trials with feedback is identical at the distribution level to that which occurs on restudy trials. Thus, study memory strength distributions after training are identical for restudied and tested items (see Fig. 1a–b, "after training").

Now consider test memory on initial test trials. When a cue is presented, *cue memory* forms, as outlined earlier. An association between cue memory and the correct answer occurs when either (1) the answer is retrieved from episodic study memory into working memory (correct trials) or (2) feedback is provided (incorrect trials). Feedback on correct test trials is assumed to have no effect on final test performance (for supporting evidence in the case of cued recall, see Pashler et al., 2005; cf. Butler, Karpicke, & Roediger, 2008, for the case of low-confidence correct responses on multiple-choice tests). Test memory strengthening in the model is not causally dependent on initial test accuracy. That assumption is reasonable given basic properties of the model; from the "perspective" of cue memory, there are simply two ways that an answer can become available from an external source (study memory or feedback), and there is no a priori reason to believe that the source in itself causally influences associative strengthening. Our assumptions that neither study nor test memory strength for tested items are causally dependent on initial test accuracy (and hence, that neither strength distribution is bifurcated by accuracy) is consistent with the hypothesis that, when there is immediate feedback, the retrieval attempt rather than retrieval success is the primary driver of learning (Kornell, Klein, & Rawson, 2015; Kornell & Vaughn, 2016; Vaughn, Hausman, & Kornell, 2016). Error learning on initial test trials is assumed to be suppressed when immediate feedback is provided (for similar conjectures, see Carrier & Pashler, 1992; Kornell & Son, 2009). We assume for simplicity that the test memory strength distribution after training is identical to the study memory strength distributions. Thus, after training, all three strength distributions across the two conditions are identical (see Fig. 1c). Strengths in study and test memory across tested items are assumed to be independent.

On the final test, correct retrieval in the restudy condition is predicted to occur for any item with a study memory strength that is above a response threshold (t), with t held constant for a given subject across all trials and phases of an experiment (see all panels of Fig. 1). In the test condition on the final test, correct retrieval may occur through study memory, test memory, or both. According to the model, the *TE* is solely dependent on the combined contribution of study and test memory; if either study or test memory were absent for a tested item, then the model would predict zero *TE* (see Fig. 1). Correct retrieval through study memory and test memory is assumed to occur all or none and independently. Finally, incorrect responses on the final test occur only when neither study nor test memory strength is above the response threshold.

Quantitative implementation

Drawing on the conceptual description above, a simple quantitative model for the ideal subject can be specified. For a randomly selected item in the restudy condition, the probability correct on the final test (P_R) is $P_R = P(S_R > t)$, where S_R is the item strength in study memory at the time of the final test and t is the response threshold. For the case of testing with feedback, the probability correct for a randomly selected item on the final test based on study memory alone is $P_{T\text{-}s} = P(S_{T\text{-}s} > t)$, and probability correct based on test memory alone is, $P_{T\text{-}t} = P(S_{T\text{-}t} > t)$, where $S_{T\text{-}s}$ is study memory strength and $S_{T\text{-}t}$ is test memory strength.

Given the properties of the model described above, probability correct through either study memory, test memory, or both is governed by the product rule for independent events,

$$P_T = P_{T-s} + P_{T-t} - P_{T-s}P_{T-t}. \tag{1}$$

Because $P_{T\text{-}s} = P_{T\text{-}t} = P_R$ in the model, Eq. 1 can be expressed as,

$$P_T = 2P_R - {P_R}^2, \tag{2}$$

and the equation for the *TE* is,

$$TE = P_T - -P_R = P_R - {P_R}^2. \tag{3}$$

The model therefore predicts that both P_T and the *TE* can be expressed solely as a function of probability correct in the restudy condition. Equation 3 is depicted by the solid-line curve in Fig. 2. Also shown (dashed line) is the logical upper bound *TE* as a function of P_R. In the model for cued recall, the maximum value of *TE* (0.25) is observed when $P_R = 0.5$. As P_R approaches 1.0 or zero, the *TE* approaches zero.

For a group of ideal subjects (an ideal experiment), the predicted grand means for P_T and *TE* are the means of the predicted subject-level P_Ts and *TE*s. Because subjects are expected to have different values of P_R, the grand mean prediction for *TE* will not correspond to a point on the model prediction curve for the ideal subject in Fig. 2 but rather will always be below that curve. That fact can be appreciated by considering a group of ideal subjects with an observed grand mean P_R of 0.5. Most or all of those subjects would be expected to have P_R values either below 0.5 or above 0.5, and in all such cases the subject-level predicted *TEs* (Eq. 3) will be less than the maximum of 0.25. Hence, a grand mean P_R of 0.5 will yield a mean *TE* of less than 0.25, with the extent less determined by the P_R distribution over subjects. That conclusion holds for all possible distributions of subject-level P_R with nonzero variance, an assertion that follows from the fact that the prediction curve in Fig. 2 becomes progressively steeper as a function of distance from $P_R = 0.5$. Thus, for *TE* data averaged over a group of subjects, the prediction curve in Fig. 2 constitutes not an exact prediction but rather the *upper bound* of an *envelope* within which *TEs* are expected to occur. The lower bound of that envelope is zero across all values of mean $P_{R.}$ That envelope constitutes exactly one third of the logically possible *TE* space. There appears to be no other discussion in the literature of either the function relating the *TE* to performance in the restudy condition or the upper bound magnitude of the *TE* for a given retrieval task and experimental design. We suggest, however, that those two characteristics of the *TE* are among the core set of phenomena to be explained.

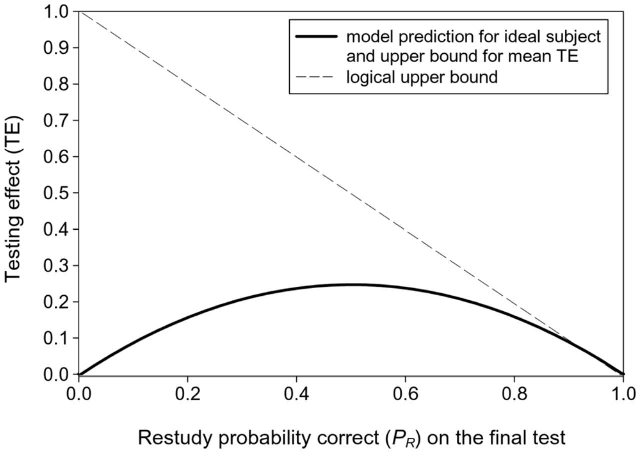

Fig. 2 The testing effect as a function of restudy probability correct. *Solid curved line*: Parameter-free dual memory model prediction for the ideal subject, and parameter-free model upper bound *TE* for (1) data averaged over subjects and (2) variants of the model in which strengths in study and test memory are positively correlated or in which test memory is weaker than the study memory. *Dashed line:* Logical upper bound *TE*. See Fig. 5 for plots of literature data

There are two additional respects in which the curve in Fig. 2 constitutes a psychologically meaningful upper bound prediction. First, it is an upper bound for alternative instantiations of the model in which the correlation between test and study memory strength over tested items is positive rather than zero (the case of independence currently assumed). A positive correlation is plausible; items that are easier to learn through study (in study memory) may also be easier to learn through testing (in test memory). Second, it is an upper bound for instantiations of the model in which the test memory strength distribution is weaker than (i.e., in Fig. 1a, shifted to the left relative to) the study memory strength distributions, a case to which we will return later.

Finally, consider data from an actual subject (in which there are a finite number of items). If the number of items in the restudy condition is large, then the equation for an unbiased estimate of the true model prediction for P_T has the same form as for the ideal subject (Eq. 2) but is expressed in terms of observed proportion correct in the restudy condition (PC_R),

$$\widehat{PC}_T = 2PC_R - {PC_R}^2, \tag{4}$$

where PC_T is the proportion correct estimate for the test condition. The unbiased estimate of the *TE* is,

$$\widehat{TE} = PC_R - {PC_R}^2. \tag{5}$$

When there are a small number of items in the restudy condition, the predictions given by Eqs. 4 and 5 will underestimate the true model prediction for P_T and TE (Eqs. 2 and 3). However, for the example case of about 20 items in the restudy condition, as in much of the TE literature, that small sample bias is negligible (about 0.01 when $P_R = 0.5$, with that amount decreasing as P_R approaches 0 or 1.0).[1]

There are two approaches to fitting the model to experimental data, both of which are used below. First, if subject-level proportion correct data are available, then quantitative predictions for PC_T and the TE for each subject that closely match the true model predictions can be calculated based on subject-level PC_R (see Eqs. 4 and 5), and those predictions can be used to predict the experiment-level mean PC_T and the mean TE. Second, if subject-level data are not available but experiment-level mean proportion correct data are, then the model can be tested by observing whether the mean TE (or its confidence interval) falls within the envelope bounded by Eq. 3 (see Fig. 2).

Implications of the model for data interpretation

For an experiment the model predicts maximum mean TE magnitude when the grand mean P_R is 0.5 and the variability in P_R across subjects are low, and a smaller magnitude as the mean P_R approaches zero or one, and (or) when P_R variability is higher. Those properties of the model may explain why *TE*s are larger in some contexts than others, without resorting to different theoretical accounts. For example, a finding that mean *TE*s for a given experiment are different for one subject population versus another (e.g., children vs. adults) might be interpreted as suggesting that retrieval practice is intrinsically less potent in one of the populations, but according to the model that conclusion would not necessarily follow. That possibility can be extended to any other experimental situation in which the TE is hypothesized to depend on the level of a second, orthogonally manipulated variable.

Empirical tests of the dual memory model

Data from our laboratory

We first tested the model against 10 data sets collected in our laboratory, all of which were originally designed to explore individual differences in, or transfer of, the *TE* (Pan, Gopal, & Rickard, 2015; Pan, Pashler, Potter, & Rickard, 2015; Pan, Wong, Potter, Mejia, & Rickard, 2016). Those data sets are the first 10 entries of Appendix A (listed under *Laboratory data*). Data from nine of the data sets had not been analyzed, nor in most cases collected, prior to model development. All 10 experiments employed similar designs and procedures, had minimum complexity, and met the earlier described criteria to which the model most directly applies. Each entailed the following:

1. Three experimental phases over two sessions: an initial study and training phase in session one, and a final test phase in session two.
2. A single presentation of each item during the initial study phase, for either 6 or 8 s per item over experiments.
3. A training phase involving random assignment of items into two subsets with counterbalanced assignment of those subsets to the restudy and testing conditions; a single presentation of each item followed by immediate feedback; equated exposure time per trial in the restudy and test conditions; and instructions to type the response on test trials as quickly and accurately as possible. Following the majority of the literature, no response was required on either initial study or restudy trials. Elements of the stimuli did not have strong preexperimental associations and thus test performance in the training phase is likely to have been mediated almost exclusively by episodic memory that was formed during the initial study phase.
4. A final test session in which each item in the restudy and testing conditions was presented once in random order for testing, using the same cued recall format as on the initial test. There was no trial time limit and no feedback was provided.

Across the 10 experiments, there was design variation in two respects that are of potential theoretical interest: (1) the retention interval between the training and final test phases (24 hrs, 48 hrs, or 1 week), and (2) the materials used (paired associate words, triple associate words, and history facts).

Results

Predictions for mean PC_T were calculated separately for each subject using Eq. 4. For each of the data sets, the number of items in the restudy condition was either 18 or 20. The item sampling bias discussed earlier when applying Eq. 4 should

[1] The item small sample bias estimates were calculated (over multiple values of P_R and restudy item sample size, n) using 100,000 simulated subjects, each with the same value of P_R and n in each simulation. The simulation for each unique combination of P_R and n can thus be understood as an *experiment* with 100,000 identical subjects. For each simulated subject, the observed proportion correct in the restudy condition on the final test was generated using n Bernoulli trials with the success parameter on each trial set to the value of P_R in the simulation. The result is a distribution of PC_R values over simulated subjects in each experiment. The simulated observed TE (assuming the parameter free model is correct) was then calculated for each subject by applying Eq. 5 to the observed PC_R. The mean of that observed TE was then compared to theoretical TE for P_R value of that simulation using Eq. 3. The difference between the simulated TE and the theoretical TE at each simulated level of P_R and n constituted the item small sample bias.

thus be negligible and was ignored. Model predictions for mean PC_T, along with observed mean PC_T and PC_R, are plotted in Fig. 3. In agreement with the model, the 95% confidence interval of the predicted PC_T contains the observed PC_T for all experiments. We also observed no systematic differences in model fit over the factors of either retention interval or material type. To explore the possibility of small but systematic prediction error, we collapsed data from all 10 experiments into a single analysis (n = 483), with results shown in the far-right-side bar graph in Fig. 3. The mean difference score (test condition predicted minus test condition observed) was 0.009, with a 95% confidence interval of ±0.02. A scatterplot of the predicted versus observed PC_T over experiments is shown in Fig. 4. As would be expected if the model is correct, the data points lie near the diagonal across the range of predicted PC_T values. Overall, the dual memory model provides good quantitative fits to these data sets, with no free parameters.

Testing the model against results in the literature

The testing effect as a function of restudy performance

We next conducted a broad literature search for *TE* experiments involving feedback on the initial test and cued recall on both the initial and final tests. Two articles that catalogued or reviewed the literature (Rawson & Dunlosky, 2011; Rowland, 2014) served as the primary sources for identifying experiments. In Rawson and Dunlosky (2011), 82 empirical articles covering a decade of retrieval practice research (2000–2010) were catalogued. In Rowland (2014), 61 empirical studies covering nearly 4 decades of research were subjected to a meta-analytic review.

We also performed a separate keyword search of the APA's PsycINFO database for peer-reviewed empirical articles using the keywords *testing effect* and *retrieval practice* and restricted to the date range (2013–2015) that was not covered by the other two sources. This search, which identified 145 candidate articles, was completed in June 2015. Combined, the published and database sources contributed 264 candidate empirical articles (excluding overlapping entries).

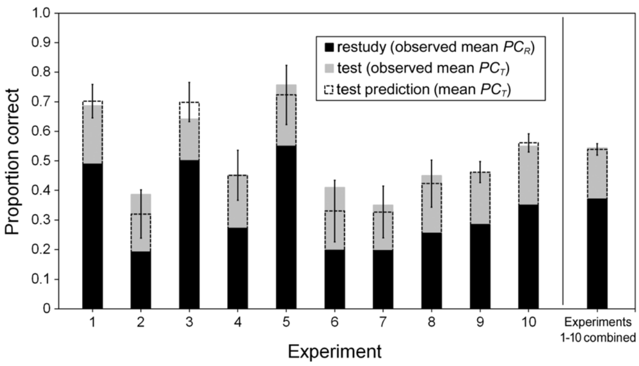

Fig. 3 Observed and predicted results for ten data sets collected in our laboratory and for all 10 data sets combined (n = 483). *Error bars* indicate 95% confidence intervals. Experiment numbers correspond to the first 10 rows of Appendix A

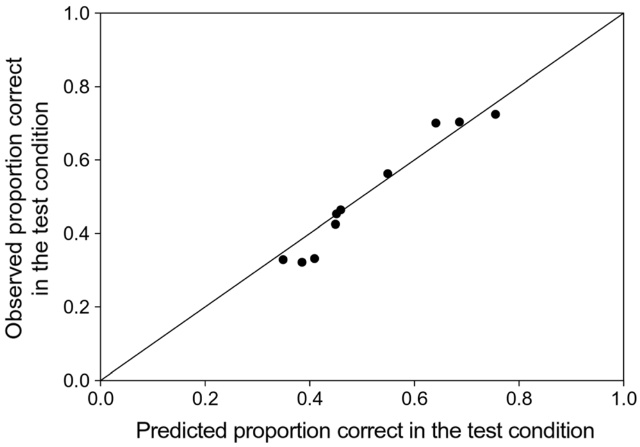

Fig. 4 Scatterplot of observed versus predicted proportion correct in the test condition for 10 data sets collected in our laboratory

In the next review phase, we examined all 264 articles to identify experiments that (a) incorporated the three-phase experiment design discussed earlier and (b) compared testing with feedback to a restudy control. The criteria for inclusion here was more lenient than the strict conditions for which the model was developed (and which held for the foregoing analyses of experiments from our laboratory). We found no experiments in the literature that met all of those conditions, thus necessitating the broader inclusion criteria. All experiments that were identified from the database search and subsequent screening process are catalogued in Appendix A (listed under *Literature search*).

For each of the 114 identified experiments, the mean *TE* and mean PC_R was recorded (see Appendix A). In cases where values were reported in graphical form but were not specified in the text, graphical pixel analysis using the technique detailed in Pan and Rickard (2015) was employed to extract mean *TE* and mean PC_R. Because subject-level data are not reported in the literature, the subject-level model predictions could not be calculated. Instead we explored whether the extracted mean *TE*s as a function of mean PC_R tend to fall within the envelope predicted by the model. If there is no pattern suggesting that the model upper bound prediction is psychologically meaningful, then the model would be refuted in this analysis. There was no expectation based on the prior literature regarding the expected outcome.

Results are shown in Fig. 5. The great majority of *TEs* are within the model prediction envelope, an envelope that constitutes one third of the *TE* space. Only five of the 114 *TE*s had confidence intervals that did not extend into the envelope (indicated by open circles), and several of those experiments had extreme design features, including extensive item repetition over multiple training sessions. Although there was insufficient reported data to calculate confidence intervals for many

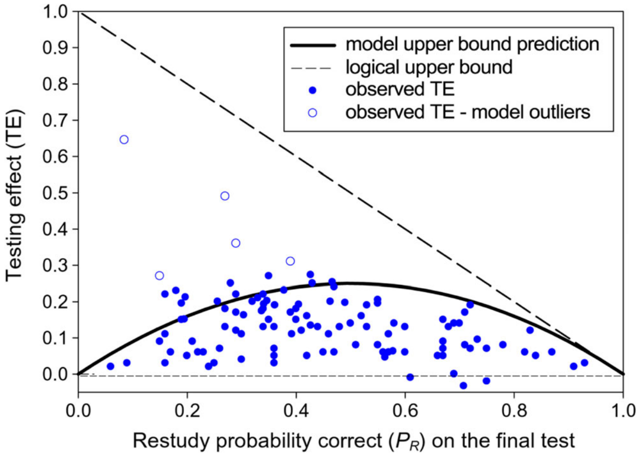

Fig. 5 Mean testing effects as a function of restudy probability correct for 114 data sets in the literature identified in our literature review

of the effects, intervals that could be calculated suggest an interval of between ±0.04 and ±0.07. It is thus not surprising from the standpoint of the model that a modest number *TE* point estimates are larger than the boundary value.

From the perspective of the parameter-free version of the model, effects that are far below the prediction upper bound and toward the center of the curve are unlikely to have been fitted well if subject-level data had been available, unless the subject-level PC_R distribution in those cases is bimodal, concentrated toward the upper and lower tails of the envelope. Those data points could be accommodated by the more general modeling framework, however, if it is assumed that, under some circumstances, test memory is weaker than study memory. As described in the next section, a number of data points in Fig. 5 are viable candidates for that possibility.

The results described above are generally consistent with the dual memory model. In particular, the upper bound prediction for the *TE* appears to be psychologically meaningful. Although the fits of the model to these data are not exact, they do constitute notable theoretical progress in our view: No other model in the literature places any constraints on the magnitude of the TE, either in absolute terms or as a function of restudy performance.

The testing retention effect for the case of feedback

In multiple studies, the magnitude of the *TE* has been demonstrated to increase as a function of retention interval for several days or weeks. That pattern can be expected to reverse at some point as natural forgetting occurs for both restudied and tested items and as proportion correct for both conditions eventually approaches zero. For the cases of testing both with and without feedback, that pattern is a natural consequence of the dual memory model. Consider first testing with feedback and the special case in which proportion correct in the restudy condition is 1.0 on a hypothetical immediate final test. The model predicts zero *TE* in that case (Eq. 3), as it must on a purely logical basis. Forgetting over time will result in decreasing proportion correct in the restudy condition and correspondingly increasing *TE* magnitude. That effect can be visualized for the ideal subject by mentally flipping Fig. 2 from left to right (such that maximum restudy performance is on the far left side of the horizontal axis) and imagining that the horizontal axis represents both restudy proportion correct and retention interval, with restudy proportion correct decreasing over increasing retention interval. If forgetting in the restudy condition were a linear function of time until P_R equals zero, then that curve would accurately reflect the shape of the model's retention prediction for the special case of perfect accuracy on an immediate final test. Wixted and Ebbesen (1991, 1997; see also Anderson & Schooler, 1991; Wickelgren, 1974; Wixted, 2004), however, have established compellingly that forgetting following study, as measured by proportion correct, follows a power function of time to a close approximation. Thus, the model prediction for *TE* as a function of retention interval when restudy proportion correct on an immediate test is 1.0 is a power function transformed version of the left-right flipped Fig. 2, such that the *TE* falls off more gradually on the right side than on the left side. That curve shape is represented by the solid line in Fig. 6, which shows predicted *TE* as a function retention interval over generic units of time. That prediction and all other retention predictions described below rest on the reasonable and simplest case assumption that power function forgetting measured in terms of proportion correct occurs at the same rate in both study and test memory.

Figure 6 also depicts the retention function for an ideal subject when restudy proportion correct on an immediate final test is 0.8, 0.5, and 0.15. In all of those cases, the predicted retention curve is a left-shifted version of the solid-line curve. For the case of 0.8, the *TE* is greater than zero on the immediate test, increases to a peak of 0.25 earlier than for the 1.0 case, and decreases more quickly than for the 1.0 case. For both the 0.5 and 0.15 cases, however, there is no increase in *TE* with increasing retention interval but rather a

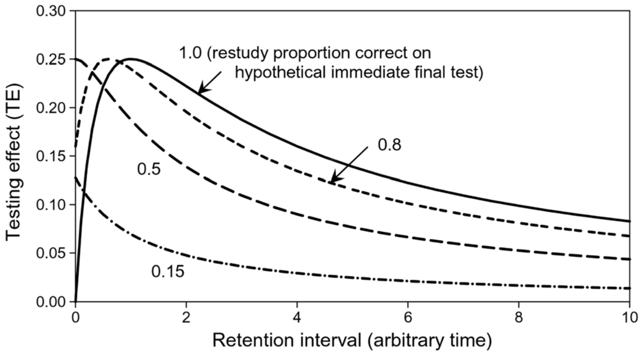

Fig. 6 Predicted testing effects as a function of retention interval and proportion correct in the restudy condition on an immediate final test. Immediate final test proportions correct of 1.0, 0.8, 0.5, and 0.15 are depicted

monotonically decreasing effect. More generally for experimental data, for any case in which restudy accuracy is above 0.5 on the shortest delay test, a range of retention intervals involving increasing *TE* magnitude is predicted, whereas for any case in which restudy accuracy is at or below 0.5 on the shortest delay test, only a decreasing *TE* magnitude as a function of retention interval is predicted.

Empirical results for testing with feedback are generally consistent with the above predictions. Multiple experiments involving cued recall and feedback exhibit positive testing effects at very short delay intervals (≤5 minutes) when restudy proportion correct is below 1.0, including Bishara and Jacoby (2008; Experiment 1); Carpenter, Pashler, Wixted, and Vul (2008); Fritz, Morris, Nolan, and Singleton (2007; Experiments 1 & 2); Jacoby, Wahlheim, and Coane (2010; Experiments 1 & 2); Morris, Fritz, and Buck (2004; Experiment 2); Rowland and DeLosh (2015; Experiment 3); and Wiklund-Hörnqvist, Jonsson, and Nyberg (2014). There is also evidence that *TEs* can increase from short- to long-delay tests when PC_R is greater than 0.5. Relative to a copy control condition that was similar to restudy, Kornell et al. (2011; Experiment 2) observed an increase in *TE* magnitude from a 2-min to 2-day retention interval. Although that result was not statistically significant, it was of similar magnitude to that expected by the dual memory model given the observed PC_R values at the two retention intervals. In the only manipulation of multiple retention intervals in the *TE* literature to date, Carpenter et al. (2008) showed that, in two experiments with high restudy accuracy (around 0.95) on the shortest (5 min) delayed final test, the *TE* increased in magnitude with increasing retention interval before decreasing. However, in their third experiment, in which restudy accuracy on the 5-min delayed final test was about 0.50, the *TE* did not significantly increase but rather appears to have only decreased with increasing delay, again consistent with model predictions.

Dual memory model fits to Carpenter, Pashler, Wixted, and Vul (2008) To gain more insight into the model's ability to account for the Carpenter et al. (2008) results, we fitted the model at the subject level for each experiment and then averaged predictions over subjects.[2] The results are shown in Fig. 7 (dashed-and-dotted lines), along with the three-parameter power function fits to the test condition data estimated by Carpenter et al. (dashed lines). The parameter-free dual memory model fitted poorly to the data from Experiment 1 but fairly well to most of the data from Experiments 2 and 3. In Experiment 2, the large increase in observed *TE* is tracked closely across the first five retention intervals (see Fig. 7d). In Experiment 3, the model predicts only decreasing *TE* with increasing retention interval (see Fig. 7f), which is roughly consistent with the test condition data (closed dots in Fig. 7e). Where the model fitted poorly across the experiments, it tended to overestimate both proportion correct in the test condition and the *TE* magnitude.

One possible explanation for that overestimation is that test memory strength was in some instances weaker in these experiments than was study memory strength. Generally, we would expect test memory strength to be more variable across different experimental designs and procedures than is study memory strength. Study memory for both the restudy and test conditions is assumed to be encoded during initial study and only strengthened during the training phase as a consequence of reactivation. That reactivation may require only modest subjective effort or executive process engagement. The degree of study memory strengthening may therefore be similar in both the restudy and test conditions. Encoding a *new* test memory during the training phase, on the other hand, may require more subjective effort and executive engagement and may require more time. If so, then experimental factors such as trial timing and subject motivation level may have more impact on the strength of test memory than on the strength of study memory.

Drawing on the above reasoning, two properties of the Carpenter et al. (2008) data raise the possibility that test memory encoding was relatively weak. First, trial timing during the training phase (for testing items: 4 s for retrieval and 1 s for feedback) was brief relative to most other studies in the literature, possibly stunting test memory formation more than the arguably more automatic study memory strengthening. Second, unlike the great majority of experiments in the literature, their subjects did not have to make an overt response on tested items in the training phase (a desired design feature given the research goals of Carpenter et al.). Hence there was no direct evidence that a response was retrieved on all test trials. There is mixed evidence regarding whether covert responding yields *TEs* that are equivalent to or smaller than those for overt responding, with the most recent work suggesting smaller *TEs* (Jönsson, Kubik, Sundqvist, Todorov, & Jonsson, 2014; Putnam & Roediger, 2013; Smith, Roediger, & Karpicke, 2013). It may also be that the effect of covert responding interacts with trial timing.

We thus explored whether a version of the model that assumes weaker test than study memory can better fit the Carpenter et al. (2008) data. We merely assumed that the proportion of items with strength above the response threshold on the final test is lower for test memory than for study memory. Otherwise, the model remained identical to the parameter-free model described earlier. To implement that single free parameter, we included a coefficient, *c*, which could take values between zero and one, beside the terms in Eq. 1 that correspond to the test memory contribution to performance, yielding P_T= $P_{T\text{-}s}$+ $cP_{T\text{-}t}$- $P_{T\text{-}s}$*$cP_{T\text{-}t}$.

[2] We thank Shana Carpenter for kindly providing the data sets for these analyses.

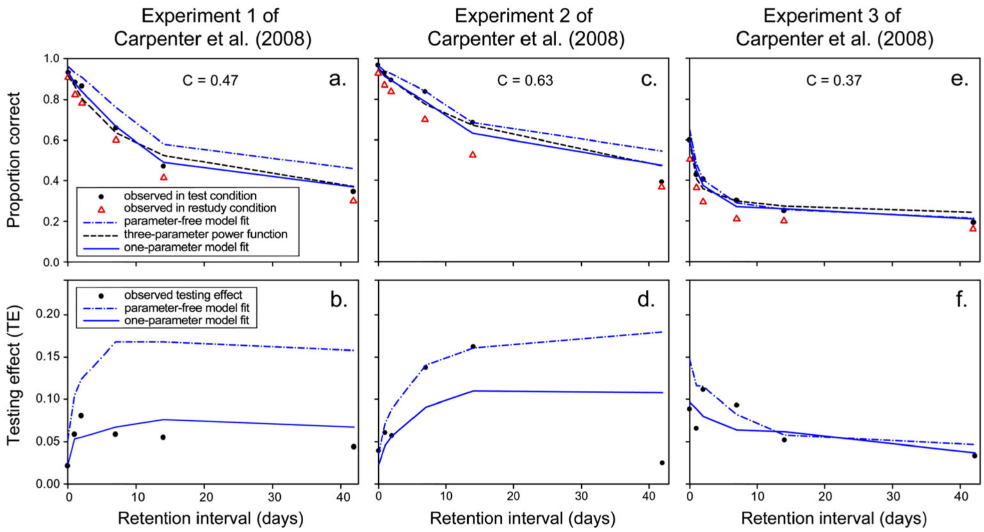

Fig. 7 Results of Experiments 1–3 of Carpenter, Wixted, Pashler, and Vul (2008). For each experiment, the best fitting three parameter power function for test condition data (as published in Carpenter et al.) as well as parameter-free and one-parameter fits of the dual memory model to both test condition and *TE* data, are shown. Data adapted with permission from Carpenter et al., *Memory & Cognition, 36*(2), p. 442.

Framed in terms of restudy performance, the equation is $P_T = P_R + cP_R - P_R{*}cP_R$.

Separately for each subject, *c* was allowed to vary as a single free parameter in the model fit, just as the three parameters of the power function were allowed to vary at the subject level in the Carpenter et al. (2008) fits. Averaged fits to the data are shown in Fig. 7 (all panels) as a solid line. The model fits are improved relative to the parameter-free case, most notably for Experiment 1. For Experiment 2, the one-parameter fit to the *TE* and PC_T may by visual analysis be poorer than that for the parameter-free case, but it is optimal by the least squares criterion (which resulted in the *TE* curve in Fig. 7d being pulled downward due to the 42-day data point).

Overall, the one-parameter model provides a reasonably good account of the Carpenter et al. (2008) data (setting aside the 42-day retention result for Experiment 2). Of note, those *one-parameter* fits are highly competitive with fits of the purely descriptive *three-parameter* power function, besting that function on a least squares basis for two of the three experiments.

Testing and the retention curve without feedback

The dual memory model as described earlier can be straightforwardly applied to the case of testing without feedback. It predicts that a correct test trial without feedback will yield the same strengthening in both study and test memory as would have been the case had there been feedback. On incorrect initial test trials without feedback, on the other hand, study memory is not accessed and hence not strengthened, and no association between cue memory and the correct response is formed. Hence, no productive associative learning occurs on those trials. Both the study and test memory strength distributions for tested items are thus bifurcated by accuracy after the training phase. That aspect of the model is consistent with work by Kornell et al. (2011) showing that testing without feedback yields bifurcated memory strengths. Given that forgetting occurs between training and final test phases, the model predicts that items incorrectly answered on an initial test with no feedback are rarely answered correctly on the final test (for empirical support, see Pashler et al., 2005).

The model's *TE* predictions for testing without feedback depend principally on two factors that do not need to be considered for the case of feedback: initial test proportion correct and the increment in strength distributions for restudied and correctly answered test items due to training. Thus, in contrast to the case of feedback, final test predictions for the case of no feedback cannot be based solely on observed final test restudy performance. Rather, model predictions were explored using simulation (for details, see Appendix B).

Model predictions for initial test proportions correct of 0.95, 0.75, and 0.50 for an ideal subject are shown in Fig. 8 as a function of retention interval. For each curve, the model predicts negative *TE*s at short delays and positive *TE*s at longer delays. The negative *TE*s at short delays become more extreme as initial test proportion correct decreases, bottoming at about -0.22 when initial test proportion correct is about 0.3 and then becoming less negative as that proportion becomes smaller. The positive testing effects become progressively smaller and shortlived as initial test proportion correct decreases below 0.5.

The prediction that the *TE* becomes larger from short to intermediate retention intervals when there is no feedback is consistent with multiple studies in the literature (for the case of cued recall, see Allen, Mahler, & Estes, 1969; Jang, Wixted, Pecher, Zeelenberg, & Huber, 2012; Kornell et al., 2011; Toppino & Cohen, 2009; for meta-analytic data, see

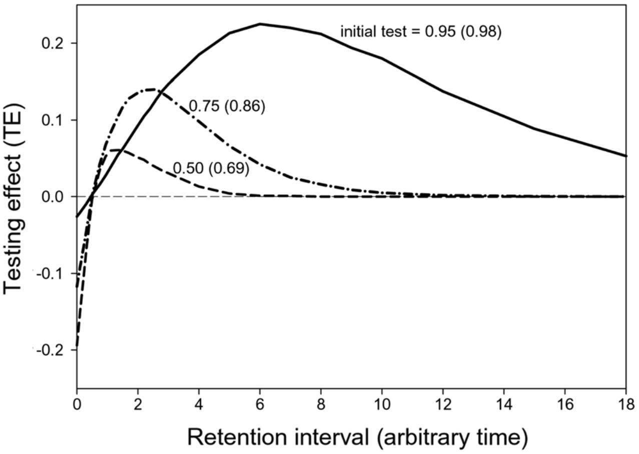

Fig. 8 Idealized subject parameter-free model predictions for the testing effect as a function of retention interval for the case of no feedback on the initial test. Initial test proportions correct of 0.95, 0.75, and 0.5 are depicted. Values in *parentheses* are corresponding predicted proportions correct in the restudy condition on a hypothetical immediate final test, as determined by the training phase strength increment value used in the simulations

Rowland, 2014). The model prediction that the *TE* can be null or negative at very short delays is also well supported in the literature (e.g., Halamish & Bjork, 2011; Jang et al., 2012; Kornell et al., 2011; Toppino & Cohen, 2009). Because there currently are no published experiments involving testing without feedback in which multiple retention intervals were manipulated, constrained fits of the model to testing without feedback retention data cannot yet be performed.

Effect of training type on later repeated testing with feedback

Across two experiments, Storm, Friedman, Murayama, and Bjork (2014) observed a theoretically important interaction between type of training and subsequent performance on a 1-week delayed final test with feedback. In their Experiment 1, the training phase involved no training after initial study (baseline), six restudy trials per item (restudy), or six testing without feedback trials per item (test), manipulated within subjects. Experiment 2 was identical, except that the testing with no feedback training condition was replaced with a testing with feedback condition. On the final test of both experiments there were six blocks, each involving one test with feedback trial per item.

In their Experiment 2 involving testing with feedback (Fig. 9b), results were as expected based on the literature: proportion correct was lowest across all six final test blocks in the baseline condition, intermediate in the restudy condition, and highest in the test condition. Of primary interest are the contrasting results of Experiment 1 (Fig. 9a). Although performance on the first final test block was slightly better in the test (with no feedback) training condition than in the restudy condition (consistent with Experiment 2), relative performance in those two conditions was flipped on Blocks 2 through 6. Thus, five prior restudy opportunities yielded a larger overall learning rate over final test blocks than did five prior testing with no feedback opportunities. Storm et al. (2014) explained their results in terms of the distribution-based bifurcation model and the new theory of disuse. Quantitative fits of their model were not reported, however.[3]

Dual memory model fits to Storm, Friedman, Murayama, and Bjork (2014) We fitted the dual memory model to the averaged data in Storm et al. (2014) to determine whether it can capture either the two-way interaction between experiment (1 vs. 2) and training condition (test vs. restudy) or the three-way interaction also involving the higher learning rates over test blocks for restudied than for tested items in Experiment 1. We used a simulation approach similar to that described earlier for the case of testing without feedback. Gamma scale parameter values for each condition on the first final test block were set to roughly yield the observed proportions correct on those blocks. In addition, for the baseline conditions, scale parameter values on all six blocks were set to roughly yield the observed proportions correct. For Blocks 2 through 6 in the restudy and test conditions, increments in scale parameter values over blocks were matched proportionally to those for the baseline condition. Model proportion correct predictions for those conditions and blocks were thus parameter free. Further details of the model fits are included in Appendix C.

For Experiment 2 (Fig. 9d), the model provided an approximate match to the Storm et al. (2014) data, demonstrating that it can roughly capture the final test repetition learning effects in the restudy and test conditions. Fits to the Experiment 1 data (Fig. 9c) show that the model also predicts both the two-way and the three-way interaction: The overall performance advantage for testing with feedback over restudy in Experiment 2 was reversed in the model fits to Experiment 1, and the crossover interaction between condition (restudy vs. test) and block in Experiment 1 was also present. The magnitude of the latter effect, however, was smaller than that observed by Storm et al.

Although the dual memory account of the Storm et al. (2014) data described here is likely too simplistic, it does demonstrate the potential of the current modeling framework to explain their critical interaction results. The fits rest mechanistically on the model predictions that (a) only items that were repeatedly answered correctly over training blocks in the

[3] In a personal communication, R. A. Bjork noted that the authors were able to fit their data quantitatively but chose not to report it.

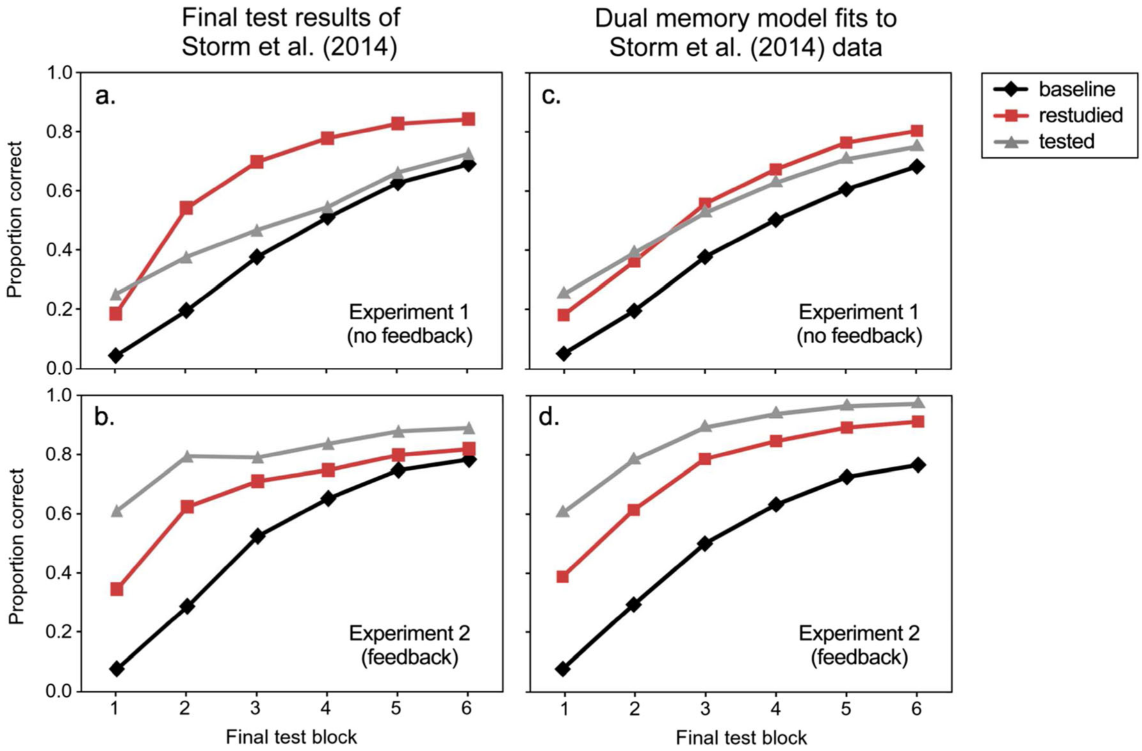

Fig. 9 Final test results of Experiments 1 and 2 of Storm, Friedman, Murayama, and Bjork (2014), and dual memory model fits to those experiments. Details of the model fits are described in Appendix C. **a** and **b** adapted with permission from Storm et al., *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*(1), p. 120.

testing without feedback condition of Experiment 1 accrued study and test memory strength during training, whereas (b) all items in the five-restudy training condition accrued study memory (but not test memory) strength during training. Hence, relative to restudied items, there was at the outset of the final test both a dual memory advantage and a strength bifurcation disadvantage for items tested without feedback. In the fits to Storm et al.'s Experiment 1 data, those two mechanisms yielded similar overall performance in those two conditions but with a larger learning rate over blocks in the restudy condition.

General discussion

We have proposed a new model of the testing effect that is based on the empirically supported tenet that retrieval practice yields a test memory that is distinct from the memory formed through initial study and that final test performance for tested items can occur through test memory, study memory, or both. From that framework, we derived a simple parameter-free model applicable to the case of cued recall. That model, and in one case a one-parameter variant of it, is unique in making successful quantitative predictions for each of the following: (a) the *TE* magnitude across 10 experiments conducted in our laboratory; (b) the *TE* magnitude as a function of restudy performance in the literature; (c) the effect of feedback; (d) the testing retention effect for the cases of both feedback and no feedback, including fits to Carpenter et al. (2008); and (e) the effects of prior learning type on subsequent learning through testing with feedback in the two experiments of Storm et al. (2014).

Assumptions of the dual memory model

Several auxiliary assumptions were made during model development. In all cases those assumptions were (a) supported by empirical results in the literature, (b) consistent with theorizing in related areas, or (c) were simplest case assumptions that did not violate empirical results in the literature. Assumptions that are consistent with prior empirical results include that (a) feedback on a *correct* test trial has negligible effect on learning (e.g., Pashler, et al, 2005; but see Butler et al., 2008, for contrasting results for low confidence correct responses using a multiple-choice test), (b) testing with feedback produces no bifurcation in either the study or the test memory strength distributions as a function of test accuracy (e.g., Kornell et al., 2015), and (c) forgetting as measured by proportion correct follows a power function of time (e.g., Wixted & Ebbesen, 1991). Assumptions corresponding to prior theoretical hypotheses include all-or-none memory retrieval, associative and retrieval independence, memory strength-threshold retrieval processes, and suppression of error associations by immediate feedback (Carrier & Pashler, 1992; Estes, 1955; Kornell et al., 2011; Ross & Bower, 1981).

Extensions to other testing effect contexts

The model developed here is restricted to the case of cued recall on both the initial and final test. The broader dual memory framework, however, can in principle be applied to any scenario in which learning through testing is compared to some nontesting task (e.g., restudy or reading). Here we illustrate that fact with examples of four relatively straightforward candidate extensions of the model.

Recognition on the initial test

In the dual memory model, a recognition test trial can reactivate and strengthen study memory just as can a restudy trial or a cued recall test trial. However, because the recognition stimulus presented on an "old" recognition trial is generally identical to an initially studied stimulus (and hence no response that is a portion of the initially studied study stimulus needs to be recalled), a separate test memory may not be encoded on recognition trials. If so, and if reactivation of study memory is largely automatic in this context, (or if subjects have roughly equivalent motivation to retrieve prior memory in the recognition and restudy conditions), then the learning difference between restudy and recognition may be minimal. The dual memory model is thus consistent with negligible or small *TEs* for the case of a training phase recognition test, and observed by Carpenter and DeLosh (2006) and Hogan and Kintsch (1971).

Transfer to stimulus-response rearranged items

We make no attempt here to extend the model fully to the diverse literature on retrieval practice and transfer of learning (for reviews, see Carpenter, 2012; Pan and Rickard, Manuscript under review). The model provides a natural candidate account, however, of recent studies in which the items presented on the final cued recall test are stimulus-response (S-R) rearrangements of items presented on the initial cued recall test. For word triplets (Pan, Wong, et al., 2016) and history and biology facts (Pan, Gopal, et al., 2015), we have shown consistently over multiple experiments that final test performance for S-R rearranged items is indistinguishable from that for restudied items and far below that of nonrearranged (i.e., tested) items. For example, if after initial study of a set of word triplets (e.g., *gift, wine, rose*), subjects are tested with feedback on retrieval of one missing word from each triplet (e.g., *gift, wine, ?*), but then on the final test are presented with S-R rearranged items from the same triplets (e.g., *wine, rose, ?*), performance on the S-R rearranged and restudied items is nearly identical (i.e., there is no obervable transfer relative to restudy).

With slight elaboration on the properties of test memory, the dual memory model provides a straightforward account of that phenomenon. As noted earlier, study memory may be best understood as a pattern learning and completion network (e.g., Bishop, 1995; Ross & Bower, 1981) that can equivalently support final test performance for any S-R arrangement, on average over items. Test memory, on the other hand, may (at least under some circumstances) allow only for retrieval in the direction of the trained cue to the trained response and thus may not support transfer to S-R rearranged triplet and fact materials. If so, then only study memory can be accessed for both restudied items and S-R rearranged test items on the final test, resulting, according to our model, in the observed equivalent performance for those sets of items. Because restudy is expected to increase memory strength and subsequent recall accuracy (e.g., Kornell et al., 2011; Pan, Wong, et al., 2016), those triplet and fact transfer results cannot be accounted for by simply assuming that none of the learning that occurs during training test trials is accessible to S-R rearranged items. Rather, the data suggest that the study memory that is strengthened during the initial test is accessible to S-R rearranged items on the final test, whereas test memory is not.

As a caveat, strong positive transfer from tested to S-R reversed paired associates has been observed (e.g., Carpenter, Pashler, & Vul, 2006), at least after one test trial per word pair (cf. Vaughn & Rawson, 2014). Paired associates appear to constitute a special case in that regard, and the mechanisms that underlie the contrasting transfer results for pairs versus triplets and facts remain to be fully understood. For further exploration and/or discussion of that issue, see Pan and Rickard (2017, Pan and Rickard, Manuscript in preparation).

Test-potentiated learning

Building on the work of Izawa (1971), Arnold and McDermott (2013) demonstrated a test-potentiated learning effect, in which five incorrect test trials without feedback prior to a study (i.e., feedback) trial yielded greater learning on that study trial—as indexed by subsequent test performance—than did one incorrect test trial without feedback prior to study. That effect was obtained in the context of no increase in proportion correct over the five incorrect test trials.

Although we do not advance a fully realized model of that phenomenon here, a feature of the dual memory model that is unique in the *TE* literature does provide a plausible mechanism for it. In the current article, it has been sufficient thus far to treat test memory in terms of a single strength value. At a finer grain size, however, two learning events are implied: encoding of cue memory and formation of an association between cue memory and the correct response, with the latter event being either concurrent with or contingent upon the former. On a later test, a correct retrieval can only occur if cue memory is reactivated *and* if the correct response is retrieved via the association.

Incorrect test trials with no feedback (whether one or five) cannot produce an association between cue memory and the correct response. Those trials can, however, enhance encoding of cue memory, the learning of which in the model requires only test stimulus presentation in the context of a retrieval task set. Five such test trials should achieve that learning to a greater extent than will one trial. Hence, during the subsequent study trial, associative strengthening between cue memory and the response can occur more effectively in the five-test condition than in the one-test condition, resulting in a greater increment in proportion correct on the next test trial (i.e., test-potentiated learning).

Via the mechanism described above, test-potentiated learning should be most potent when formation of cue memory is relatively difficult. That condition appears to hold in the Arnold and McDermott (2013) study, in which materials were paired associates involving unfamiliar, nonnative language Russian cue words and native language English response words (e.g., *medved–bear*). Further, assuming that presentation of an unfamiliar foreign word as a cue is not likely to spark rich semantic activation, an alternative account of the potentiation phenomenon based on semantic processes (e.g., Hays, Kornell, & Bjork, 2013) would not seem applicable in this case.

The foregoing discussion makes it clear (or at least highly plausible) that a separate, episodic cue memory can form on incorrect training phase test trials with no feedback. In our view it is not a far stretch to conclude that, as the dual memory model predicts, separate cue memory also forms on correct test trials and on incorrect test trials with feedback.

Free recall on the final test

There is a relatively straightforward extension of the model to the case of cued recall on the initial test and free recall on the final test (e.g., Carpenter, 2009; Carpenter & DeLosh, 2006; Carpenter, Pashler, et al., 2006; Fritz et al., 2007; Halamish & Bjork, 2011; Karpicke & Zaromb, 2010; Pan, Rubin, & Rickard, 2015; Peterson & Mulligan, 2013; Rowland, 2014; Rowland & DeLosh, 2015). As reviewed by Hintzman (2016), sequential presentation of paired associate items (as in cued recall training in the *TE* paradigm) does not appear to yield measurable interitem associations, provided that task instructions or other task properties do not lead subjects to believe that interitem associations are important to learn. Thus, final test free recall following cued recall training can be plausibly modeled by assuming independent memories for each item encountered during training. For items in the restudy condition, final test performance can be mediated by study memory only, whereas for items in the test condition, retrieval can occur through study memory, test memory, or both, resulting in greater recall probability in the test condition. That conclusion from the model likely holds regardless of the mechanistic details of free recall on the final test. It remains to be seen, however, whether that simple extension of the model can provide good quantitative fits.

Halamish and Bjork (2011; Experiment 2) explored the case of cued recall on the initial test and (most pertinently here) cued versus free recall on the final test. Their design involved *no feedback* on the initial test and a brief (8.4 min) delay between the training and final test phase. As both the dual memory and the distribution-based bifurcation models can predict, a negative *TE* was observed for cued recall on the final test. Of most interest, however, a *positive TE* was observed for free recall on the same final test. Halamish and Bjork explained those contrasting results in terms of their distribution of associations bifurcation model and the fact that free recall is typically more difficult than is cued recall. In their account, the extra difficulty of free recall has functionally the same effect as would a longer delay period (essentially requiring a higher strength threshold for correct responding), hence resulting in the shift to a positive testing effect for free recall. Because the dual memory model also incorporates the bifurcation effect for testing without feedback, it can explain their result in the same way and with the same level of specificity (see Fig. 8 for the shift from negative to positive *TE* as retention interval and plausibly other factors that affect retrieval difficulty, increases).

Further tests of the dual memory model

Because the model makes quantitative predictions with limited flexibility, there are straightforward ways to further evaluate it. One approach is to more densely sample the space of experimental manipulations to which the model was applied in this manuscript, including joint manipulations of task (such as restudy, testing with feedback, and testing without feedback) and retention interval. A second approach would be to test the core premise of the model experimentally, using manipulations that could be hypothesized to moderate the relative influence of study and test memory on the final test, or to disassociate cue memory learning and learning of the association between cue memory and the response. Three examples of such manipulations have already been discussed: recognition versus cued recall on the initial test, tested versus S-R rearranged items on the final test, and test-potentiated learning.

Another approach would be to compare the model predictions to those of alternative quantitative models that can successfully explain a similar set of *TE* phenomena. That approach awaits development of an alternative model that meets that criterion. Analogously, fair evaluation of the assumptions in the dual memory model as well as the free parameters that it may require for data fitting can be made only in comparison to an empirically competitive model of quantitative aspects of *TE* phenomena.

Relations to other theories

The dual memory model appears to be the first in the *TE* literature that applies to the cases of both feedback and no feedback. Some researchers have suggested that *TE* theorizing is most profitably restricted to the case of no feedback (e.g., Karpicke et al., 2014). We have taken the opposite approach, first developing the model for the case of feedback and then applying it straightforwardly to the case of no feedback. That approach may also be productive in the exploration of alternative theories, wherein any complication of including feedback in initial model development may be more than offset by the benefit of a more integrative account.

As noted earlier, the dual memory model incorporates study memory strength bifurcation as a function of accuracy for items tested *without feedback*, a property of testing that was first identified by Kornell et al. (2011) and Halamish and Bjork (2011). Our model's prediction of a negative *TE* on an immediate final test for that case thus relies on the same mechanism as does the distribution-based bifurcation model that was proposed by the same authors. Those models differ, however, in their hypotheses about the cause of the positive-going (and subsequently decreasing) *TE* as retention interval increases (see Fig. 8). The bifurcation model attributes that effect to the greater memory strength for tested than for restudied items along a single dimension, but it does not provide a mechanistic account of that strength difference as the dual memory model does. Further, unlike the bifurcation model as developed to date in the literature, the dual memory model applies to the cases of both feedback and no feedback, among several other phenomena, and it has been implemented and tested quantitatively.

Several theories in which a task-specific or unique processes are thought to occur on test trials, including but not limited to more elaborative retrieval (Carpenter, 2009), greater mediator effectiveness (Pyc & Rawson, 2010), or more effective neural network error correction (Mozer et al., 2004) could potentially be integrated with the dual memory model as a mechanism that moderates test memory strength but not (or less so) study memory strength. According to our model, however, the *TE* should still be observed when the possibility of semantic elaboration or the use of effective mediators appears to be minimized. As Karpicke et al. (2014) has noted, several studies that meet those criteria have yielded the expected testing effects (Carpenter & DeLosh, 2005; Carpenter & Kelly, 2012; Coppens, Verkoeijen, & Rikers, 2011; Kang, 2010).

The dual memory model may also prove useful as a reference case in future theory development. Assumptions of separate and independent processes are common in quantitative theory development, both within psychology (e.g., stimulus sampling theory; Estes, 1955; see also Atkinson & Shiffrin, 1968; Hintzman, 2010) and more generally. Because such models can make precise, constrained, and testable predictions—and because they often posit nonambiguous mechanisms—they can play an important role in ruling out some subclasses of models in favor of others (e.g., Benjamin & Tullis, 2010; Snodgrass & Townsend, 1980).

Conclusions

The dual memory model provides a well-supported quantitative account of multiple phenomena in the testing effect literature, with empirical coverage that is well beyond that provided by other models to date. The model draws attention to the potentially central role that new episodic encoding during test trial cue presentation may play in the testing effect, and it identifies a candidate mathematical relation between restudy performance and the testing effect that may be useful in future theory development. Despite the successes of the model, however, it is not advanced here as a comprehensive account of all testing effect phenomena. It currently provides no account of the effects of free recall on the initial test, certain cases of transfer, semantic relatedness effects, and immediate versus delayed feedback, among other phenomena. Given the variety of contexts in which retrieval practice enhances learning, we suspect that two or more distinct mechanisms may underlie the testing effect.

Author note Thanks to Bijan Malaklou for assistance with data extraction. S. C. Pan is supported by a National Science Foundation (NSF) Graduate Research Fellowship.

Appendix A

Testing with feedback studies included in tests of the dual memory model

Source	Reference	Condition	Design		Proportion correct		
			N	Delay	Testing	restudy	*TE*
Laboratory data[1]							
	Pan, Gopal, et al., 2015	Exp 1, AP History facts	38	48 h	0.35	0.20	0.15
	Pan, Pashler, et al., 2015	Exp 1, paired associate words	120	24 h	0.46	0.29	0.17
		Exp 2, paired associate words	122	24 h	0.55	0.36	0.19
	Pan, Wong, et al., 2016	Exp 1, triple associate words	42	7 d	0.45	0.26	0.20
	Pan and Rickard, Manuscript in preparation	Exp 1, paired associate words	33	24 h	0.69	0.49	0.20
		Exp 1, paired associate words	25	7 d	0.39	0.19	0.20
		Exp 1, triple associate words	32	24 h	0.64	0.50	0.14
		Exp 1, triple associate words	29	7 d	0.45	0.27	0.18
	Pan and Rickard, Unpublished manuscript	Exp 1, paired associate words	42	24 h	0.76	0.55	0.20
		Exp 1, paired associate words	42	7 d	0.41	0.20	0.21
Literature search[2]							
	Baghdady et al., 2014	Diagnostic accuracy, immediate	55/57	0 h	0.74	0.67	0.07
		Feature list, immediate	55/57	0 h	0.73	0.75	-0.02
		Diagnostic accuracy, delayed	55/57	7 d	0.72	0.67	0.05
		Feature list, delayed	55/57	7 d	0.60	0.61	-0.01
	Barcroft, 2007	Posttest 1	24	0 m	0.61	0.56	0.05
	Bishara & Jacoby, 2008	Exp 1, young adults	18	3 m	0.95	0.83	0.12
		Exp 1, older adults	18	3 m	0.82	0.67	0.15
	Brewer & Unsworth, 2012	Paired-associate testing task	107	24 h	0.51	0.45	0.06
	Butler, 2010	Exp 1a, factual, same test	48	24 h	0.76	0.27	0.49
		Exp 1a, conceptual, same test	48	24 h	0.70	0.39	0.31
	Carpenter, Pashler, et al., 2006	Exp 1	43	18-24 h	0.72	0.58	0.14
		Exp 2	19	18-24 h	0.64	0.50	0.14
	Carpenter et al., 2008	Exp 1, 5 min delay	55	5 m	0.93	0.91	0.02
		Exp 1, 1 day delay	55	24 h	0.88	0.82	0.06
		Exp 1, 2 day delay	55	48 h	0.86	0.78	0.08
		Exp 1, 7 day delay	55	7 d	0.66	0.60	0.06
		Exp 1, 14 day delay	55	14 d	0.47	0.42	0.05
		Exp 1, 42 day delay	55	42 d	0.34	0.30	0.04
		Exp 2, 5 min delay	57	5 m	0.96	0.93	0.03
		Exp 2, 1 day delay	57	24 h	0.93	0.87	0.06
		Exp 2, 2 day delay	57	48 h	0.89	0.84	0.05
		Exp 2, 7 day delay	57	7 d	0.84	0.70	0.14
		Exp 2, 14 day delay	57	14 d	0.68	0.52	0.16
		Exp 2, 42 day delay	57	42 d	0.39	0.36	0.03
		Exp 3, 5 min delay	44	5 m	0.59	0.51	0.08
		Exp 3, 1 day delay	44	24 h	0.43	0.36	0.07
		Exp 3, 2 day delay	44	48 h	0.41	0.30	0.11
		Exp 3, 7 day delay	44	7 d	0.30	0.21	0.09

	Exp 3, 14 day delay	44	14 d	0.25	0.20	0.05
	Exp 3, 42 day delay	44	42 d	0.19	0.16	0.03
Carpenter et al., 2009	Immediate review group	37	270 d	0.08	0.06	0.02
	Delayed review group	38	270 d	0.12	0.09	0.03
Carrier & Pashler, 1992	Exp 4	60	2 m	0.71	0.66	0.05
Coane, 2013	Young, Study-Study vs. Study-Test	24/21	10 m	0.69	0.69	0.00
	Old, Study-Study vs. Study-Test	20/21	10 m	0.62	0.56	0.06
Finley et al., 2011	Exp 2, test vs. restudy	80	10 m	0.58	0.40	0.18
Fritz et al., 2007	Exp 1, testing vs. rehearsal	15/15	3 m	0.72	0.47	0.25
	Exp 2, testing vs. restudy	10/10	24 h	0.75	0.55	0.20
Goossens, Camp, Verkoeijen, & Tabbers, 2014	Exp 1, restudy vs. RP	39/41	7 d	0.82	0.75	0.07
	Exp 2, restudy vs. RP	42/40	7 d	0.83	0.69	0.14
Goossens, Camp, Verkoeijen, Tabbers, & Zwaan, 2014	Pairs, SSSSSSS vs. SSSSTST	30	7 d	0.53	0.41	0.12
	Story, SSSSSSS vs. SSSSTST	30	7 d	0.41	0.36	0.05
Jacoby et al., 2010	Exp 1, SSSSSS vs. STTTTT	40	0 m	0.79	0.71	0.08
	Exp 2, SSSSSS vs. STTTTT	36	5 m	0.73	0.60	0.13
Kang, 2010	Exp 1	33/33	10 m	0.66	0.53	0.13
	Exp 2	39/39	24 h	0.56	0.39	0.17
	Exp 3	60	10 m	0.63	0.55	0.08
Kang & Pashler, 2014	Exp 1, low value, 2 trials	38	48 h	0.26	0.24	0.02
	Exp 1, low value, 4 trials	38	48 h	0.50	0.35	0.15
	Exp 1, high value, 2 trials	38	48 h	0.28	0.22	0.06
	Exp 1, high value, 4 trials	38	48 h	0.49	0.36	0.13
	Exp 2, no incentive, 2 trials	59	48 h	0.29	0.23	0.06
	Exp 2, no incentive, 4 trials	59	48 h	0.52	0.34	0.18
	Exp 2, incentive, 2 trials	59	48 h	0.23	0.17	0.06
	Exp 2, incentive, 2 trials	59	48 h	0.23	0.17	0.06
	Exp 2, incentive, 4 trials	59	48 h	0.51	0.29	0.22
	Exp 3, no incentive, 2 trials	120	48 h	0.24	0.15	0.09
	Exp 3, no incentive, 4 trials	120	48 h	0.45	0.27	0.18
	Exp 3, incentive, 2 trials	120	48 h	0.28	0.25	0.03
	Exp 3, incentive, 4 trials	120	48 h	0.55	0.40	0.15
Kang et al., 2007	Exp 2, short answer training	48	72 h	0.57	0.46	0.11
Kang et al., 2013	Exp 1, comprehension	41	0 h	0.63	0.57	0.06
	Exp 1, production	41	0 h	0.40	0.27	0.13
	Exp 2, comprehension	59	0 h	0.57	0.44	0.13
	Exp 2, production	59	0 h	0.34	0.19	0.15
Karpicke & Blunt, 2011	Exp 1, repeated study vs. RP	20/20	7 d	0.66	0.46	0.20
Keresztes et al., 2014	Short retention interval group	13	20 m	0.67	0.71	-0.03
	Long retention interval group	13	7 d	0.50	0.39	0.11
Kornell & Son, 2009	Exp 1, feedback condition	19	5 m	0.53	0.47	0.06
Kornell et al., 2011	Exp 2, copy vs. T + FB condition	40	10 m	0.91	0.72	0.19
	Exp 2, copy vs. T + FB condition	40	48 h	0.71	0.47	0.24
Kromann et al., 2009	Control vs. intervention group	40/41	14 d	0.83	0.73	0.10
LaPorte & Voss, 1975	Exp 1, restudy vs. Qs + answers	24/24	7 d	0.56	0.43	0.13
Larsen et al., 2013	Study vs. written test	41	180 d	0.61	0.48	0.13
Lipko-Speed et al., 2014	Exp 1, study vs. test + feedback	27	48 h	0.33	0.26	0.07
	Exp 2, study vs. test + feedback	30	48 h	0.27	0.16	0.11
McDermott et al, 2014	Exp 3, restudy vs. SA quiz	116	60 h	0.81	0.68	0.13
Metcalfe et al., 2007	Exp 1, computer vs. self-study	14	7 d	0.73	0.09	0.65
	Exp 2, computer vs. self-study	18	7 d	0.41	0.29	0.12

	Exp 3, computer vs. self-study	13	7 d	0.76	0.67	0.09
Morris & Fritz, 2000	Exp 2, simple game vs. repetition	100+	30 m	0.65	0.29	0.36
Morris & Fritz, 2002	Original game vs. no retrieval	88/64	30 m	0.72	0.53	0.19
Morris et al., 2004	Exp 2, semantic instructions	12/12	5 m	0.38	0.16	0.22
	Exp 2, no semantic instructions	12/12	5 m	0.53	0.28	0.25
Pan, Gopal, et al., 2015	Exp 2, AP Biology facts	58	48 h	0.61	0.38	0.23
	Exp 4, AP History facts, 1 term	45	24 h	0.47	0.30	0.16
	Exp 4, AP History facts, 2 terms	45	24 h	0.51	0.33	0.17
	Exp 4, AP History facts, 3 terms	45	24 h	0.53	0.34	0.19
Pan, Wong, et al. 2016	Exp 2, triple associate words	58	7 d	0.55	0.35	0.20
	Exp 3, triple associate words	60	7 d	0.70	0.43	0.27
Peterson & Mulligan, 2013	Exp 2	28/28	5 m	0.79	0.72	0.07
Putnam & Roediger, 2013	Exp 2, restudy vs. type + aloud	50	48 h	0.68	0.43	0.25
	Exp 2, restudy vs. overt	25	48 h	0.52	0.32	0.20
Pyc & Rawson, 2010	Cue-only retrieval group	20/20	7 d	0.42	0.15	0.27
	Cue + mediator given group	20/20	7 d	0.47	0.34	0.13
	Cue + mediator recall group	20/20	7 d	0.41	0.18	0.23
Rohrer et al., 2010	Exp 1, standard test	28	24 h	0.56	0.34	0.22
	Exp 2, standard test	28	24 h	0.58	0.42	0.16
Rowland & DeLosh, 2015	Exp 4, 30 s condition	18	0.5 m	0.60	0.41	0.19
	Exp 4, 90 s condition	20	1.5 m	0.54	0.33	0.21
Storm et al., 2014	Exp 2, first final delayed test	18	7 d	0.62	0.35	0.27
Sumowski et al., 2010	Healthy controls, spaced	16	45 m	0.68	0.57	0.11
Wartenweiler, 2011	Exp 1	32	1 h	0.64	0.58	0.06
Wiklund-Hörnqvist et al., 2014	SS vs. ST_{fb}	43/40	5 m	0.88	0.71	0.17

Note. Exp = Experiment; RP = retrieval practice; *TE* = testing effect. *N* indicates the number of subjects in the experiment in total, or if it was a between-subjects design, the number to the left of the slash is the total in the testing condition, while the number to the right of the slash is the total in the restudy condition. [1]Laboratory data = 10 data sets initially used to test the dual memory model; [2]Literature search = papers gathered from Rawson and Dunlosky (2011), Rowland (2014), and database searches that were used in subsequent tests of the dual memory model

Appendix B

Simulated predictions for testing without feedback as a function of retention interval

The simulation implements the parameter-free model, adapted to the case of no feedback as described in the main text. Each data point that forms the prediction lines in Fig. 8 was based on 100,000 random samples from a gamma distribution of item strengths for an ideal subject. The gamma distribution shape parameter was set to 2.0 for all simulations, and the response threshold, *t*, was set to 1.0. Post study phase memory strengths were set to achieve initial test proportion correct of 0.95, 0.75, and 0.50. To yield those initial test values, the corresponding gamma distribution scale parameters were set to 2.85, 1.04, and 0.6. To model training phase learning in the restudy condition and for correctly answered test items, the scale parameter for each distribution after initial study was multiplied by 1.5. For simulated incorrect (below threshold) initial test trials, the study memory strength distribution was not incremented. Hence, those items were never answered correctly on the simulated final test. Reduced memory strength (S_1) over a retention interval for each simulated item was modeled using a power function, $S_1 = S_0 * (\textit{retention interval} + 1)^{-1}$, where S_0 is the item strength immediately after the training phase.

Appendix C

Simulations of the Storm, Friedman, Murayama, and Bjork (2014) data

The Storm, Friedman, Murayama, and Bjork (2014) data were simulated using the same framework that was used to model the retention function for the case of testing without feedback. However, in this case, only final test performance was directly modeled. The gamma distribution shape parameter was held

Table 1 Table of gamma scale parameter values (as a function of experiment, condition, memory type, and final test block)

			Final Test Block					
Source	Training condition	Memory type	1	2	3	4	5	6
Experiment 1								
	Restudy	Test memory	—	1.53	2.03	2.38	2.54	2.61
		Study memory	**2.00**	2.58	3.41	4.09	4.77	5.23
	Test Correct	Test memory	**4.50**	5.80	7.67	9.20	10.74	11.76
		Study memory	**4.50**	5.80	7.67	9.20	10.74	11.76
	Test Incorrect	Test memory	—	1.53	2.03	2.43	2.83	3.10
		Study memory	—	1.53	2.03	2.43	2.83	3.10
	Baseline	Test memory	—	1.53	2.03	2.43	2.83	3.10
		Study memory	**1.32**	**1.70**	**2.25**	**2.70**	**3.15**	**3.45**
Experiment 2								
	Restudy	Test memory	—	1.80	2.38	2.88	3.33	3.60
		Study memory	**3.00**	4.05	5.37	6.49	7.50	8.11
	Test	Test memory	**2.90**	3.19	5.19	6.27	7.25	7.84
		Study memory	**2.90**	3.19	5.19	6.27	7.25	7.84
	Baseline	Test memory	—	1.80	2.38	2.88	3.33	3.60
		Study memory	**1.48**	**2.00**	**2.65**	**3.20**	**3.70**	**4.00**

Note, Boldface values correspond to data points in Storm et al. (2014) that were used to fix the values of gamma distribution scale parameters in the simulations

constant at 2.0, and the response threshold, t, was held constant at 6.25. Separate simulations were performed for Experiments 1 and 2. To mimic the combined effects of learning during the training phase and forgetting during the retention interval, gamma scale parameters (β) for the first final test block in all conditions was set such that predicted proportion correct roughly approximated observed proportion correct, incorporating the parameter-free model assumption of identical strength distributions in study and test memory for the test conditions. Because the rate of learning with repeated testing is not specified in the model, β values for the baseline condition were set such that predicted proportion correct in that condition roughly matched observed proportion correct across all six final set blocks. Those preset values, shown in bold in Table 1, were then used to fully constrain the rate of increase in β from block to block in the restudy and test conditions, as detailed below. Thus, the model makes parameter-free predictions for observed proportion correct from Blocks 2 through 6 for both the test and restudy conditions. These simulations were performed for the ideal subject, whereas the Storm et al. data are averaged over experimental subjects. Because we were not attempting to optimize quantitative fits in this case but rather are exploring whether the interactions in the Storm et al. data can be produced by the model, that fact is not consequential.

Consider the baseline condition of Experiment 1 (see Table 1). On the first final test block there had been no prior test trials in that condition, so there is no test memory (as represented by the dashes in the corresponding table cell); β for study memory was set to 1.32 to yield the proportion correct observed by Storm et al. (0.05) on that block. On Block 2, we assume that the scale parameter for test memory (which is first formed on final test Block 1) is 90% that of study memory on Block 2; because study memory strength was weak on the first final test block in the baseline conditions (as indexed by low proportion correct in that case), it was reasonable to assume that β for study memory strength on the second test block would be only slightly greater than β for test memory strength on that block. No other percent alternatives were searched in the simulations. The scale parameter for study memory on Block 2 was then set such that the predicted proportion correct matched the observed data, yielding study memory β value of 1.70 and test memory β of 1.53. For Blocks 3 through 6, the same parameter setting procedure was repeated, with the test memory β value always incremented by the same proportion as the study memory β value.

Because test memory strength is independent of study memory strength in the model, β on each final block within each experiment is identical for all cases in which there were no correct test trials during training, including the baseline and restudy conditions, and incorrectly answered items in the test condition of Experiment 1 (see Table 1). For incorrectly answered test items in Experiment 1, both study and test memory are assumed to be absent or inaccessible on the first final test block. For simplicity, the identical distributions assumptions for study and test memory in the parameter-free model are assumed to hold throughout the final test, both for simulated items that were correctly and incorrectly answered during the training phase. Based on results of Storm et al. (2014), 30% of the simulated tested items in Experiment 1 were modeled as have been correctly answered during training and 70% were modeled as having been answered incorrectly. Procedures for setting β values for Experiment 2 were identical to those for Experiment 1, with the exception that, because Experiment 2 involved testing with feedback, the β values for correctly and incorrectly answered items were modeled identically.

References

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*(4), 463–470.

Anderson, J. R., & Schooler, L. J. (1991). Reflections of the environment in memory. *Psychological Science, 2*(6), 396–408.

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

Atkinson, R. C., & Raugh, M. R. (1975). An application of the mnemonic keyword method to the acquisition of a Russian vocabulary. *Journal of Experimental Psychology: Human Learning and Memory, 1*(2), 126–133. doi:10.1037/0278-7393.1.2.126

Atkinson, R. C., & Shiffrin, R. M. (1968). Human memory: A proposed system and its control processes. In K. W. Spence & J. T. Spence (Eds.), *The psychology of learning and motivation* (Vol. 2, pp. 89–195). New York, NY: Academic Press.

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

Bajic, D., & Rickard, T. C. (2009). The temporal dynamics of strategy execution in cognitive skill learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35*(1), 113–121.

Barcroft, J. (2007). Effects of opportunities for word retrieval during second language vocabulary learning. *Language Learning, 57*(1), 35–56. doi:10.1111/j.1467-9922.2007.00398.x

Benjamin, A. S., & Tullis, J. (2010). What makes distributed practice effective? *Cognitive Psychology, 61*(3), 228–247.

Bishara, A. J., & Jacoby, L. L. (2008). Aging, spaced retrieval, and inflexible memory performance. *Psychonomic Bulletin & Review, 15*(1), 52–57. doi:10.3758/PBR.15.1.52

Bishop, C. M. (1995). *Neural networks for pattern recognition*. Oxford, UK: Oxford University Press.

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. (1994). Memory and metamemory considerations in the training of human beings. In J. Metcalfe & A. P. Shimamura (Eds.), *Metacognition: Knowing about knowing* (pp. 185–205). Cambridge, MA: MIT Press.

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*(1), 32–41. doi:10.1016/j.jml.2011.02.005

Brewer, G. A., & Unsworth, N. (2012). Individual differences in the effects of retrieval from long-term memory. *Journal of Memory and Language, 66*(3), 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*(5), 1118–1133. doi:10.1037/a0019902

Butler, A. C., Karpicke, J. D., & Roediger, H. L. (2008). Correcting a meta-cognitive error: Feedback increases retention of low-confidence correct responses. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34*(4), 918–928. doi:10.1037/0278-7393.34.4.918

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

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

Carpenter, S. K., & DeLosh, E. L. (2005). Application of the testing and spacing effects to name learning. *Applied Cognitive Psychology, 19*(5), 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*(2), 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*(3), 443–448. doi:10.3758/s13423-012-0221-2

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

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*(6), 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*(5), 826–830. doi:10.3758/BF03194004

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

Coane, J. H. (2013). Retrieval practice and elaborative encoding benefit memory in younger and older adults. *Journal of Applied Research in Memory and Cognition, 2*(2), 95–100. doi:10.1016/j.jarmac.2013.04.001

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*(3), 351–357. doi:10.1080/20445911.2011.507188

Crutcher, R. J., & Ericsson, K. A. (2000). The role of mediators in memory retrieval as a function of practice: Controlled mediation to direct access. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26*(5), 1297–1317. doi:10.1037/0278-7393.26.5.1297

Delaney, P. F., Reder, L. M., Staszewski, J. J., & Ritter, F. E. (1998). The strategy-specific nature of improvement: The power law applies by strategy within task. *Psychological Science, 9,* 1–7.

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

Estes, W. K. (1955). Statistical theory of distributional phenomena in learning. *Psychological Review, 62*(5), 369–377.

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(4), 289–298. doi:10.1016/j.jml.2011.01.006

Fritz, C. O., Morris, P. E., Nolan, D., & Singleton, J. (2007). Expanding retrieval practice: An effective aid to preschool children's learning. *The Quarterly Journal of Experimental Psychology, 60*(7), 991–1004. doi:10.1080/17470210600823595

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

Glover, J. A. (1989). The "testing" phenomenon: Not gone but nearly forgotten. *Journal of Educational Psychology, 81*(3), 392399. doi:10.1037/0022-0663.81.3.392

Goossens, N. A. M. C., Camp, G., Verkoeijen, P. P. J. L., Tabbers, H. K., & Zwaan, R. A. (2014). The benefit of retrieval practice over elaborative restudy in primary school vocabulary learning. *Journal of Applied Research in Memory and Cognition, 3*(3), 177182. doi:10.1016/j.jarmac.2014.05.003

Goossens, N. A. M. C., Camp, G., Verkoeijen, P. P. J. L., & Tabbers, H. K. (2014). The effect of retrieval practice in primary school vocabulary learning. *Applied Cognitive Psychology, 28*(1), 135–142. doi:10.1002/acp.2956

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

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

Hintzman, D. L. (2010). How does repetition affect memory? Evidence from judgments of recency. *Memory & Cognition, 38*(1), 102–115.

Hintzman, D. L. (2016). Is memory organized by temporal contiguity? *Memory & Cognition, 44,* 365–375. doi:10.3758/s13421-015-0573-8

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

Izawa, C. (1971). The test trial potentiating model. *Journal of Mathematical Psychology, 8*(2), 200–224.

Jacoby, L. L., Wahlheim, C. N., & Coane, J. H. (2010). Test-enhanced learning of natural concepts: Effects on recognition memory, classification, and metacognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36*(6), 1441–1451. doi: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: The role of retrievability. *The Quarterly Journal of Experimental Psychology, 65*(5), 962–975. doi:10.1080/17470218.2011.638079

Jönsson, F. U., Kubik, V., Sundqvist, M. L., Todorov, I., & Jonsson, B. (2014). How crucial is the response format for the testing effect? *Psychological Research, 78*(5), 623–633. doi:10.1007/s00426-013-0522-8

Kang, S. H. K. (2010). Enhancing visuospatial learning: The benefit of retrieval practice. *Memory & Cognition, 38*(8), 1009–1017. doi:10.3758/MC.38.8.1009

Kang, S. H. K., Gollan, T. H., & Pashler, H. (2013). Don't just repeat after me: Retrieval practice is better than imitation for foreign vocabulary learning. *Psychonomic Bulletin & Review, 20*(6), 1259–1265. doi:10.3758/s13423-013-0450-z

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

Kang, S. H. K., & Pashler, H. (2014). Is the benefit of retrieval practice modulated by motivation? *Journal of Applied Research in Memory and Cognition, 3*(3), 183–188. doi:10.1016/j.jarmac.2014.05.006

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

Karpicke, J. D., Lehman, M., & Aue, W. R. (2014). Retrieval-based learning: An episodic context account. In *The Psychology of Learning and Motivation* (Vol. 61, pp. 237–284).

Karpicke, J. D., & Zaromb, F. M. (2010). Retrieval mode distinguishes the testing effect from the generation effect. *Journal of Memory and Language, 62*(3), 227–239. doi:10.1016/j.jml.2009.11.010

Keresztes, A., Kaiser, D., Kovács, G., & Racsmány, M. (2014). Testing promotes long-term learning via stabilizing activation patterns in a large network of brain areas. *Cerebral Cortex, 24*(11), 3025–3035. doi:10.1093/cercor/bht158

Kole, J. A., & Healy, A. F. (2013). Is retrieval mediated after repeated testing? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 39*(2), 462–472. doi:10.1037/a0028880

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

Kornell, N., Klein, P. J., & Rawson, K. A. (2015). Retrieval attempts enhance learning, but retrieval success (versus failure) does not matter. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 41*(1), 283–294. doi:10.1037/a0037850

Kornell, N., & Son, L. K. (2009). Learners' choices and beliefs about self-testing. *Memory, 17*(5), 493–501. doi:10.1080/09658210902832915

Kornell, N., & Vaughn, K. E. (2016). How retrieval attempts affect learning: A review and synthesis. In B. H. Ross (Ed.), *Psychology of learning and motivation* (Vol. 65, pp. 183–212). Amsterdam, The Netherlands: Elsevier.

Kromann, C. B., Jensen, M. L., & Ringsted, C. (2009). The effect of testing on skills learning. *Medical Education, 43*(1), 21–27. doi:10.1111/j.1365-2923.2008.03245.x

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

Larsen, D. P., Butler, A. C., Lawson, A. L., & Roediger, H. L. (2013). The importance of seeing the patient: Test-enhanced learning with standardized patients and written tests improves clinical application of knowledge. *Advances in Health Sciences Education, 18*(3), 409–425. doi:10.1007/s10459-012-9379-7

Lipko-Speed, A., Dunlosky, J., & Rawson, K. A. (2014). Does testing with feedback help grade-school children learn key concepts in science? *Journal of Applied Research in Memory and Cognition, 3*(3), 171–176. doi:10.1016/j.jarmac.2014.04.002

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

Metcalfe, J., Kornell, N., & Son, L. K. (2007). A cognitive-science based programme to enhance study efficacy in a high and low risk setting. *European Journal of Cognitive Psychology, 19*(4/5), 743–768. doi:10.1080/09541440701326063

Morris, P. E., & Fritz, C. O. (2000). The name game: Using retrieval practice to improve the learning of names. *Journal of Experimental Psychology: Applied, 6*(2), 124–129. doi:10.1037/1076-898X.6.2.124

Morris, P. E., & Fritz, C. O. (2002). The improved name game: Better use of expanding retrieval practice. *Memory, 10*(4), 259–266. doi:10.1080/09658210143000371

Morris, P. E., Fritz, C. O., & Buck, S. (2004). The name game: Acceptability, bonus information and group size. *Applied Cognitive Psychology, 18*(1), 89–104. doi:10.1002/acp.948

Mozer, M. C., Howe, M., & Pashler, H. (2004). Using testing to enhance learning: A comparison of two hypotheses. In K. Forbus, D. Gentner, & T. Regier (Eds.), *Proceedings of the twenty-sixth annual conference of the cognitive science society* (pp. 975–980). Mahwah, NJ: Erlbaum.

Newell, A., & Rosenbloom, P. S. (1981). Mechanisms of skill acquisition and the law of practice. In J. R. Anderson (Ed.), *Cognitive skills and their acquisition* (pp. 1–55). Hillsdale, NJ: Erlbaum.

Pan, S. C., Gopal, A., & Rickard, T. C. (2015). Testing with feedback yields potent, but piecewise, learning of history and biology facts. *Journal of Educational Psychology, 107*(4). doi:10.1037/edu0000074

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

Pan, S. C., Rubin, B. R., & Rickard, T. C. (2015b). Does testing with feedback improve adult spelling skills relative to copying and reading?. *Journal of Experimental Psychology: Applied, 21*(4), 356–369. doi:10.1037/xap0000062

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

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

Pan, S. C., Wong, C. M., Potter, Z. E., Mejia, J., & Rickard, T. C. (2016). Does test-enhanced learning transfer for triple associates? *Memory & Cognition 44*(1). doi:10.3758/s13421-015-0547-x

Pashler, H., Cepeda, N. J., Wixted, J. T., & Rohrer, D. (2005). When does feedback facilitate learning of words? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31*(1), 3–8. doi:10.1037/0278-7393.31.1.3

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

Putnam, A. L., & Roediger, H. L. (2013). Does response mode affect amount recalled or the magnitude of the testing effect? *Memory & Cognition, 41*(1), 36–48. doi:10.3758/s13421-012-0245-x

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

Pyc, M. A., & Rawson, K. A. (2012). Why is test–restudy practice beneficial for memory? An evaluation of the mediator shift hypothesis. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 38*(3), 737–746. doi:10.1037/a0026166

Raugh, M. R., & Atkinson, R. C. (1975). A mnemonic method for learning a second-language vocabulary. *Journal of Educational Psychology, 67*(1), 1–16. doi:10.1037/h0078665

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

Rickard, T. C. (1997). Bending the power law: A CMPL theory of strategy shifts and the automatization of cognitive skills. *Journal of Experimental Psychology: General, 126*(3), 288–311. doi:10.1037/0096-3445.126.3.288

Rickard, T. C. (2007). Forgetting and learning potentiation: Dual consequences of between-session delays in cognitive skill learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33*(2), 297–304. doi:10.1037/0278-7393.33.2.297

Rickard, T. C., & Bajic, D. (2006). Cued recall from image and sentence memory: A shift from episodic to identical elements representation. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 32*(4), 734–748. doi:10.1037/0278-7393.32.4.734

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

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

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

Ross, B. H., & Bower, G. H. (1981). Comparisons of models of associative recall. *Memory & Cognition, 9*(1), 1–16.

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

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

Smith, M. A., Roediger, H. L., & Karpicke, J. D. (2013). Covert retrieval practice benefits retention as much as overt retrieval practice. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 39*(6), 1712–1725. doi:10.1037/a0033569

Snodgrass, J. G., & Townsend, J. T. (1980). Comparing parallel and serial models: Theory and implementation. *Journal of Experimental Psychology: Human Perception and Performance, 6*(2), 330–354.

Storm, B. C., Friedman, M. C., Murayama, K., & Bjork, R. A. (2014). On the transfer of prior tests or study events to subsequent study. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 40*(1), 115–124. doi:10.1037/a0034252

Sumowski, J. F., Chiaravalloti, N., & DeLuca, J. (2010). Retrieval practice improves memory in multiple sclerosis: Clinical application of the testing effect. *Neuropsychology, 24*(2), 267–272. doi:10.1037/a0017533

Thios, S. J., & D'Agostino, P. R. (1976). Effects of repetition as a function of study-phase retrieval. *Journal of Verbal Learning and Verbal Behavior, 15*(5), 529–536.

Toppino, T. C., & Cohen, M. S. (2009). The testing effect and the retention interval: Questions and answers. *Experimental Psychology, 56*(4), 252–257. doi:10.1027/1618-3169.56.4.252

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

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

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*(1), 74–93.

Wickelgren, W. A. (1974). Single-trace fragility theory of memory dynamics. *Memory & Cognition, 2*(4), 775–780.

Wiklund-Hörnqvist, C., Jonsson, B., & Nyberg, L. (2014). Strengthening concept learning by repeated testing. *Scandinavian Journal of Psychology, 55*(1), 10–16. doi:10.1111/sjop.12093

Wixted, J. T. (2004). On common ground: Jost's (1897) law of forgetting and Ribot's (1881) law of retrograde amnesia. *Psychological Review, 111*(4), 864–879. doi:10.1037/0033-295X.111.4.864

Wixted, J. T. (2007). Dual-process theory and signal-detection theory of recognition memory. *Psychological Review, 114*(1), 152–176. doi:10.1037/0033-295X.114.1.152

Wixted, J. T., & Ebbesen, E. B. (1991). On the form of forgetting. *Psychological Science, 2*(6), 409–415.

Wixted, J. T., & Ebbesen, E. B. (1997). Genuine power curves in forgetting: A quantitative analysis of individual subject forgetting functions. *Memory & Cognition, 25*(5), 731–739.