Journal of Memory and Language 134 (2024) 104476

Contents lists available at ScienceDirect

Journal of Memory and Language

journal homepage: www.elsevier.com/locate/jml

Interaction between the testing and forward testing effects in the case of Cued-Recall: Implications for Theory, individual difference Studies, and application☆

Mohan W. Gupta [a], Steven C. Pan [b], Timothy C. Rickard [a,*]

[a] University of California, San Diego, United States
[b] National University of Singapore, Singapore

ARTICLE INFO

Keywords:
Cued recall
Testing effect
Retrieval practice
Forward testing effect
Dual-memory model
Individual differences

ABSTRACT

Recall from episodic memory has been shown to enhance both memory for the retrieved information (e.g., relative to a restudy control condition; the *testing effect, or* TE) and memory for different, subsequently studied materials (the *forward testing effect, or* FTE). Hence, the TE may be subject to an FTE confound when training in a TE experiment involves either testing prior to restudy or when restudied and tested items are randomly mixed. Across two cued-recall TE experiments, we show that (1) a potent FTE confound exists in the test-first but not the mixed training design, (2) there are no other evident learning related interactions between restudied and tested items across three frequently used training phase task orderings, and (3) the predictions of the dual-memory model of test-enhanced learning – which posits that a test trial creates a memory that is separate from the initially encoded study memory, yielding two routes to retrieval for tested items – are held both when there is and is not a confounding FTE. Further, our results yielded no evidence for two accounts of the FTE (the proactive interference and reset of encoding hypotheses) as applied to cued recall but are consistent with two alternative accounts (the strategy change and increasing effort hypotheses). Through distribution analyses we identify a novel and potent FTE individual differences effect that can be accommodated by the latter accounts. Finally, we show that at least three large-*n* studies exploring individual differences in the TE are confounded by the FTE, compromising conclusions in those papers about the efficacy of the TE across individuals.

Introduction

The mechanisms by which retrieval enhances episodic memory have been investigated extensively over the last two decades, in large part using the *testing effect* (TE) and the *forward testing effect* (FTE) paradigms. The TE is the finding that, following initial study for a set of items, subsequent learning and retention is greater if they are tested rather than restudied. That result is particularly robust if correct answer feedback (henceforth, feedback) is provided after each test trial. The typical TE experiment for the case of cued recall – the type of testing that we explore here – involves a three-phase design. Participants first study a set of items, such as facts or word pairs (the *study phase*). In the subsequent *training phase*, task type is usually manipulated within-participant (about 71% of studies based on the review in Pan el., 2018; see also Appendix A); half of the items are restudied, and half are tested (either with or without feedback across different experiments). Following a retention interval ranging from a few minutes to several weeks, a *final test* phase is administered, wherein there is a cued recall test trial for each item that was restudied (the final test *restudy condition*) or tested (the *test condition*). In their review of the cued recall literature, Rickard and Pan (2018) observed a TE in 96% of published experiments, i.e., final test proportion correct in the test condition (PC_T) was nearly always larger than final test proportion correct in the restudy condition (PC_R).

Having established the TE, researchers have turned to related topics, including transfer of test-based learning (for a review, see Pan & Rickard, 2018), application (e.g., McDaniel et al., 2007; Schwieren, et al., 2017), theoretical accounts (e.g., Carpenter & DeLosh, 2006; Kornell et al., 2011; Karpicke et al., 2014; Rowland, 2014; Endres & Renkl,

☆ We thank our multiple research assistants, including Max Bishop, Muzi Chen, Charles Dupont, Andrew Nguyen, and Gayathri Sridhar, without whom the empirical work from our laboratory would not have been possible.

* Corresponding author at: Department of Psychology, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA, United States
E-mail address: trickard@ucsd.edu (T.C. Rickard).

https://doi.org/10.1016/j.jml.2023.104476
Received 7 July 2023; Received in revised form 22 September 2023; Accepted 31 October 2023

2015; Hopper & Huber, 2018; Rickard & Pan, 2018), and individual difference effects (Brewer & Unsworth, 2012; Tse & Pu, 2012; Agarwal, Finley, Rose, & Roediger, 2017; Minear et al. 2018; Robey, 2019).

The FTE, in contrast, occurs when testing potentiates subsequent study-based learning for a different set of items (Cho et al., 2017; Yang et al., 2018a; Yang et al., 2021;). The FTE, which is often referred to as test-potentiated new learning, has typically been studied using list-based materials (Szpunar et al., 2008; Cho et al., 2017; Pastötter et al., 2017; Yang et al., 2017; Yang et al., 2018a). The standard design involves participants learning five-word lists. Half of the participants are given a free recall test after studying each list, followed by a cumulative free recall test. The remaining half study the first four lists without being tested, are tested after studying the fifth list, and are then given the same cumulative test. The list five test and the list five items on the cumulative test are the key comparisons between groups. The group tested on all five lists performs better on both measures than the group tested only after having studied list five. Hence, prior testing facilitates learning of new word lists via study. That phenomenon constitutes the FTE, an effect that has also been also observed for cued-recall (Soderstrom & Bjork, 2014; Sohlberg et al., 2021) and multiple choice tests (Choi et al., 2020).

Multiple mechanisms of the FTE have been proposed, as summarized in Sohlberg et al. (2021): *proactive interference* (Szpunar et al., 2008), *reset of encoding* (Pastötter et al., 2011), *change in strategy* (Soderstrom & Bjork, 2014; Chan et al., 2018; Cho et al, 2017), *improved accessibility* (Wissman et al., 2011), *test expectancy* (Weinstein et al., 2014), and *increased effort* (Cho et al., 2017). Four of those accounts are most applicable to the current work on testing through cued recall (proactive interference, reset of encoding, change in strategy, and increased effort). The proactive interference hypothesis posits that interference builds up during successive restudy trials. However, testing between restudy bouts creates a new context with which the restudied items can be associated, reducing that interference (Sohlberg et al., 2021; Dang et al., 2021 for age differences; cf. Wissman et al., 2011). The similar reset of encoding hypothesis states that the memory encoding through study decreases in efficiency over consecutive trials and that testing reestablishes this efficiency through contextual segregation of the learned information, yielding the FTE. The change in strategy hypothesis encompasses a couple of mechanisms. First, a test may encourage participants to spend more time studying versus if there was no preceding test (Soderstrom & Bjork, 2014), and (or) the use of more effective encoding strategies (Yang et al., 2017). Second, in the case of list learning, testing may encourage semantic and temporal clustering of subsequently studied material (Yang et al., 2018b; Chan et al., 2018; see Dang et al., 2021 for age related differences). Finally, the increased effort account states that failed memory recall can motivate greater effort during subsequent study (Cho et al., 2017).

A possible forward testing effect confound in the TE literature

One primary goal of this paper is to explore whether there is an FTE confound in the three-phase cued-recall TE paradigm that we summarized earlier, and if this confound depends on the training phase design. Consider three training designs in the literature that differ only in training task order: *test-first* blocked training (wherein all test items are presented prior to restudy items), *restudy-first* blocked training, and randomly *mixed* training (Fig. 1). For the restudy-first design, there cannot be an FTE confound as defined in the literature. For the test-first design, there may well be an FTE confound in the form of larger PC_R (and consequently a smaller TE) than the restudy-first group. For the mixed training design, it is also possible that test trials create an FTE confound because in that design many test trials proceed many restudy trials. Abel and Roediger (2017) found no difference in the TE between mixed training and blocked training (i.e., test-first and restudy first combined, counterbalanced over participants), but they did not perform separate analysis of the test-first and restudy-first data. Thus, we cannot make strong inferences from that study regarding the possibility of an FTE confound in either the test-first or mixed cases.

Although the possibility of an FTE confound has not been previously explored for cued-recall, it has recently been demonstrated for free recall of word lists. Mulligan et al. (2022) showed that repeated recall of words from one studied list can substantially increase learning that occurs during restudy of a subsequent list with different words. They discuss protection of proactive interference and enhanced encoding as plausible theoretical bases for their FTE results. The Mulligan et al. (2022) results suggest that an FTE confound may also exist for cued-recall. However, the FTE for cued-recall has received much less attention in the literature, and its applicability and potency in the current experiments is uncertain.

Implications of an FTE confound for the cued-recall TE

Establishing whether and when the FTE confound exists in cued-recall TE experiments is valuable for at least three reasons. First, large FTE confounds could lead to substantial attenuation of TE magnitude, in turn underestimating the efficacy of testing as a tool for enhancing learning. As reported in Appendix A, we explored the potential prevalence of FTE confounds by analyzing the training phase designs of 56

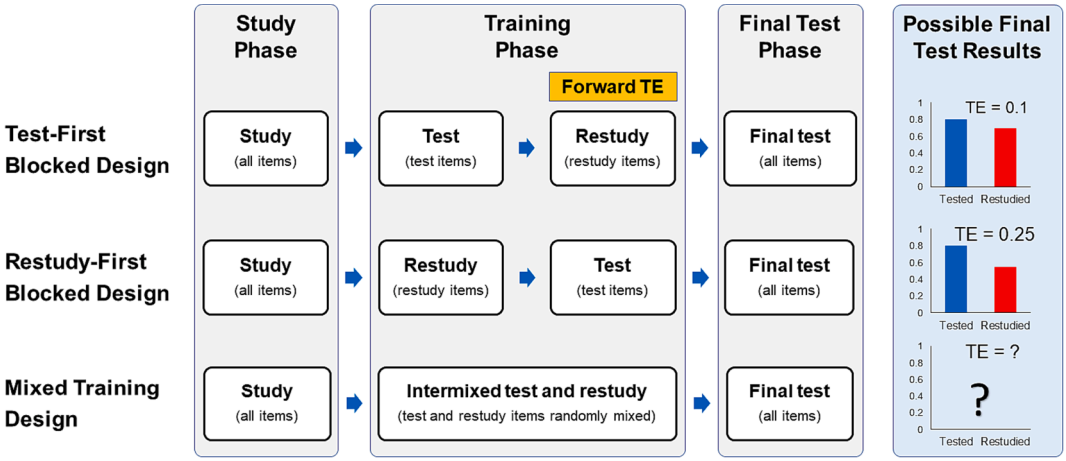

Fig. 1. A diagram showing the three training phase designs. All the designs are the same in the study and final test phases. In the test-first training design, items are first trained through testing and then separate items are trained through restudy. We hypothesize that this arrangement creates an FTE confound. In the restudy-first design, items are first trained through restudy and then separate items are trained through restudy. This design cannot create an FTE confound. In the mixed-training design, items trained through restudy test are randomly intermixed. This design may or may not create an FTE confound, although the dual-memory model assumes that it does not.

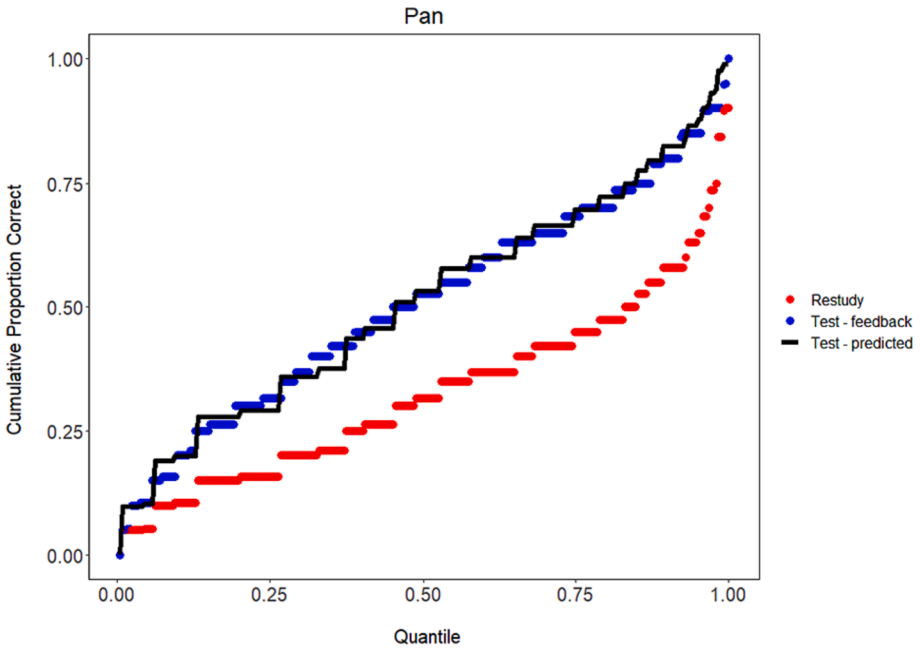

Fig. 2. Cumulative distribution graph of the combined data from Experiments 1 and 2 of Pan et al. (2015). The Test - predicted line is from the dual memory model prediction (Equation 4).

published experiments involving cued recall testing, across 41 papers. In 70% of those experiments, the training design is potentially subject to that confound, involving either mixed training (36%) or blocked training with counterbalancing such that half of the participants received test-first training (34%). Second, efforts to develop quantitative models of the TE – such as the dual-memory (Rickard & Pan, 2018) model that in part motivated the current work – would be compromised by an unrecognized FTE confound. Third, efforts to establish individual difference factors in the TE, along with potential translation of those findings into educational contexts, could be seriously compromised by an unrecognized FTE confound. In a later section of the current paper, we report evidence for FTE confounds in three such studies.

The Dual-Memory model of test-enhanced learning

A second primary goal of this paper is to determine whether the dual-memory model of test-enhanced learning can account for the TE both when an FTE confound is present and when it is not. The theoretical framework and the corresponding quantitative model are described in detail in Rickard and Pan (for application of the model to transfer of test enhanced learning, see Rickard and Pan, 2020; for application to the effect that the amount of prior study has on the TE, see Gupta et al., 2021). Brief descriptions of the conceptual framework and the assumptions that underlie the quantitative model are provided in Appendix B.

The core claim of the model is that, whereas restudy strengths the memory retrieval route that is formed during initial study (study memory), a test with feedback trial both strengthens study memory and forms a separate memory of the test event (test memory), yielding a separate test memory route to retrieval on the final test. Those two routes to retrieval yield higher probability correct on the final test for tested items than for restudied items. The quantitative model predicts that the expected proportion correct in the test condition (PC_T-expected) of the final test for a given participant (within participant experiments) is an approximate quadratic function of the observed proportion correct in the restudy condition for that same participant:

$$PC_T - expected = 2PC_R - PC_R^2 \quad (1)$$

The expected TE according to the model is:

$$TE - expected = \left(2PC_R - PC_R^2\right) - PC_R = PC_R - PC_R^2 \quad (2)$$

The model thus makes PC_T predictions separately for each participant, with no free parameters. Although a model with no free parameters surely has boundaries in its applicability, it has proven useful as a simplest case quantitative implementation of the dual-memory framework when applied to the cued-recall TE design that is used in the current paper and frequently in the literature.

Prior model fits to proportion correct means and distributions

Rickard and Pan (2018) demonstrated close fits of the dual-memory model to the mean PC_T (and hence, the *TE*) across different material types (paired associates, triplets, and history facts) and retention intervals (1, 2, and 7 days). They also showed that the model provides a viable account of the *TE* for the cases of both feedback and no feedback during training, as well as the *TE* retention function for the cases of both feedback and no feedback.

Of particular relevance to the current paper, Rickard (2020) used cumulative distribution plots to show that the model's predictions hold not just for means but also for the proportion correct distribution. An example cumulative plot and model fit to data from Pan et al. (2015; Experiments 1 and 2 combined) is shown in Fig. 2 (n = 201). In that plot, PC_R is sorted from smallest to largest across all participants. There are 201 equally space quantile values on the x-axis, scaled from zero to one, representing each of the 201 participants. The PC_T values are independently sorted from smallest to largest. Hence, in this type of plot, a given participant's PC_R and PC_T values would occur on the same quantile only by chance. Finally, the *PC_T-predicted* value at each quantile is the Equation (1) model prediction, based on the PC_R value at the same quantile (for further discussion, see Rickard, 2020).

Implications of an FTE confound for the model

The dual memory model assumes implicitly that learning and retrieval occur independently for items in the test and restudy conditions. For current purposes, it assumes that there is no FTE confound influencing PC_R. Hence, use of Equation (1) to predict PC_T for an experiment that has an FTE confound would not yield a valid test of the

model. Rather, the appropriate design involves two randomized groups in which one group does not have an FTE confound (e.g., restudy-first training) and the other group does (e.g., test-first training). In that design, the model prediction for PC_T based on PC_R from the restudy-first group is expected to hold for both that group and for the test-first group.

The presence of an FTE confound in the case of mixed training would constitute a more fundamental problem for the model. All of the TE experiments conducted by the authors of this manuscript to date, and against which the model has been tested, have involved mixed training. We have assumed, implicitly, that mixed training does not produce an FTE confound. But if that confound *is* present in the mixed case, then the dual-memory model, designed to account for only the TE, would have fitted well in our experiments to PC_T results that are actually determined by both the TE and the FTE.

Data Availability

All data, analyses, and stimuli are available at https://osf.io/5xafs/. For further requests please email the corresponding author Dr. Tim Rickard, trickard@ucsd.edu.

Overview of the experiments

In Experiment 1, we compared two blocked training groups: restudy-first and test-first. Our hypothesis was a larger PC_R and a smaller TE in the test-first group due to an FTE confound, but no group difference in PC_T. As noted earlier, we also expected good model fits to PC_T for both groups, when using PC_R for the restudy-first group to make both predictions. Results were consistent with those expectations. In Experiment 2, we again included restudy-first and test-first training groups, along with two additional groups: mixed training and restudy-only training. That experiment replicated Experiment 1, demonstrated no FTE confound in the mixed training case, and allowed us to rule out both the proactive interference and reset of encoding hypothesis as major drivers of the FTE as applied to cued recall, at least for the materials used in the current experiments. The Experiment 2 results also suggest that, aside from the FTE confound, there are no effects of training phase task order on final test performance in either the restudy or test conditions, in line with the assumption of our model.

Experiment 1: Restudy-first vs. Test-first blocked training

The experimental design is nearly identical in most respects to our prior experiments on this topic (e.g., Pan et al., 2015), the primary exception being that one group involved restudy-first training and the other group test-first training.

Methods

Participants

One hundred forty-two undergraduate students participated and were compensated with course credit. Data from six participants was excluded due to non-completion of the second session of the experiment. Among the 136 remaining participants, 64 had been randomly assigned to the restudy-first group, and the remaining 72 had been randomly assigned to the test-first group. Those sample sizes per group exceed that of most published TE experiments.

Materials

The materials were 80 weakly associated English word pairs that were drawn from the Nelson, et al. (1998) Free Association Norms database (a subset of which has been used in our prior work, e.g., Pan et al., 2015). Words were between 4 and 7 letters in length. Mean forward associate strengths for the word pairs was 0.028. Forward and backwards associative strengths were similar (*mean difference* = 0.0084).

Design and procedure

The three-phase testing effect experimental design referenced at the outset of this article was used. In session 1, participants completed the study phase wherein they studied a series of word pairs, one word-pair per trial. They next completed a training phase wherein half of the word pairs were trained via restudy and the remainder via cued recall testing with feedback. Assignment of word pairs to restudy or testing was counterbalanced across participants and each word pair was trained once using either restudy or testing. Crucially, during the training phase, restudy and testing occurred in separate blocks, and the order in which those blocks occurred (i.e., restudy-first or test-first) was assigned randomly between participants. The final test phase was conducted 48 h later.

Study phase. Participants read instructions stating that they were to learn a series of word pairs and to pay close attention to each pair. They then viewed each pair once serially for 6 s, and in random order. Each pair appeared at the center of a computer screen in boldface, size 40, serif font.

Training phase. For each participant, 40 of the studied word pairs were randomly assigned to the restudy condition and 40 to the test condition. Depending on the random group assignment, participants completed the restudy block first or the test block first. The restudy block began with instructions stating that participants were to study again the word pairs they had just previously studied. The presentation of each restudied pair was identical to that in the study phase (6 s per trial), in a random order determined anew for each participant. The test block began with instructions stating that participants were to be presented with one word per pair and asked to type the missing word (which was represented by "???"); once 5 s had elapsed, they were to attend to the correct answer feedback (which would appear on the computer screen for 1 s, replacing the "???"). Presentation of each word and the "???" per tested pair involved nearly identical on-screen positioning, the same font characteristics, and same total trial time (6 s) as in the restudy block. The order of word pairs in the test block were randomized anew for each participant.

Final test phase. A cued-recall test was administered for each of the 80 words pairs, across eighty trials. For pairs from the training phase test condition, the same response was required. There was no feedback and trials were self-paced. Word pairs were tested in random order.

Scoring and analysis

In both experiments, Participants' typed responses were scored as correct only if they matched exactly with the correct answer word. Analyses in both experiments involved t-tests and Analysis of Variance, along with Bayes factors for critical tests. In accordance with the suggested default Cauchy prior value (Wagenmakers et al., 2018), a uniform r (prior) value of 0.707 was employed across all experimental assessments. The Bayes factor was evaluated employing Raftery's criteria (Raftery, 1995) for interpretation, whereby Bayes factors falling within the range of 1–3 were deemed weak evidence, 3–20 indicated positive evidence, 20–150 denoted strong evidence, and Bayes factors exceeding 150 were considered to provide very strong evidence. BF_{10} indicates in favor of the alternative and BF_{01} indicates in favor of the null.

Results and discussion

Training phase means

As expected, mean proportion correct on tests trials was statistically indistinguishable for the restudy-first (.30) and test-first (.34) groups, $t(134) = 1.37$, $p = .17$.

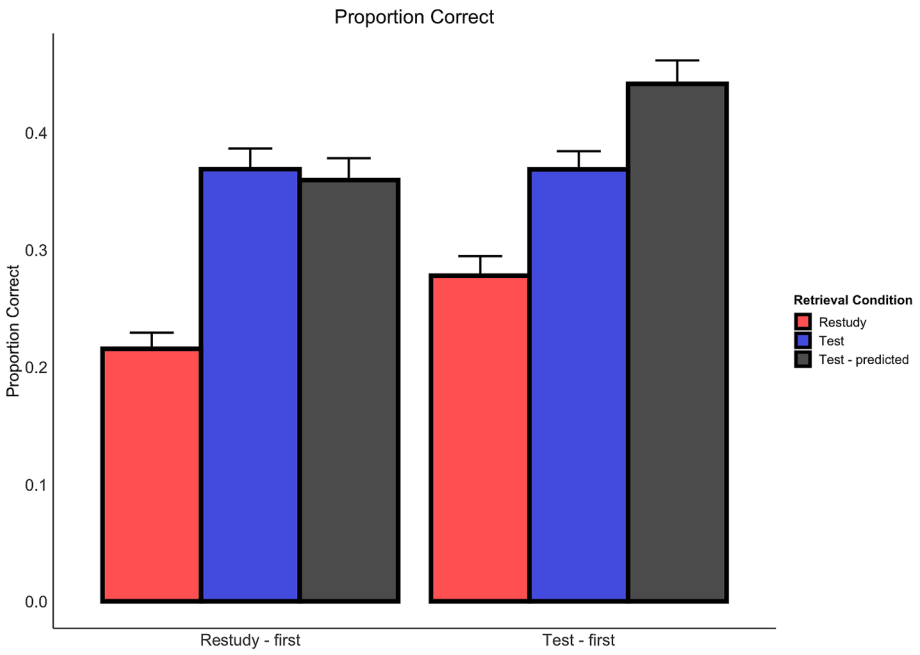

Fig. 3. Mean observed restudy (PC_R) and test (PC_T) condition proportions correct for the Experiment, along with the model prediction for the test conditions, separately for the restudy-first group and the test-first group. Given the FTE, the appropriate test of the model involves comparison of the predicted PC_T for the restudy-first group with observed PC_T for the test-first group. Error bars are standard errors calculated separately for each condition and group.

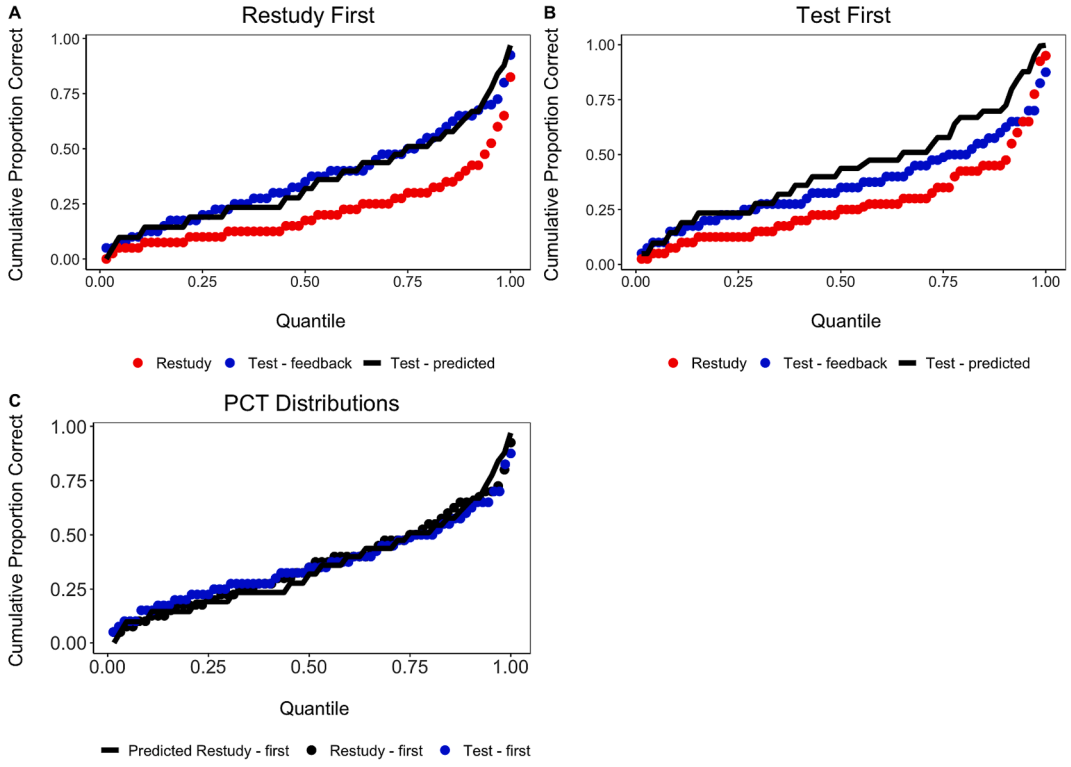

Fig. 4. Cumulative distribution model fits to the final test data from the Experiment. Panel a: Restudy-first group; Panel b: test-first group. Panel c: PC_T for both groups along with the predicted-PC_T for the restudy-first group.

Final test phase means

Mean proportions correct in the restudy and test conditions for both groups are shown in Fig. 3, along with model predictions which will be discussed in the next section. A 2 (Training Order; restudy-first vs. test-first; between participants) by 2 (Task, restudied vs. tested; within participants) factorial ANOVA was performed on the participant-level proportion correct. There was no statistically significant main effect of Training Order, $F(1, 134) = 1.04$, $p = .31$, $\eta_p^2 = 0.008$, $BF_{01} = 3.32$. There were, however, significant effects of both Task, $F(1, 134) = 159.4$, $p < .0001$, $\eta_p^2 = 0.543$, $BF_{10} = 51940$, indicating a robust *TE*, and the Training Order by Task interaction, $F(1, 134) = 10.76$, $p = .0013$, $\eta_p^2 = 0.074$, $BF_{10} = 7474$. Because the mean PC_T was virtually identical for the two training orders (.36 and .37), the observed interaction primarily

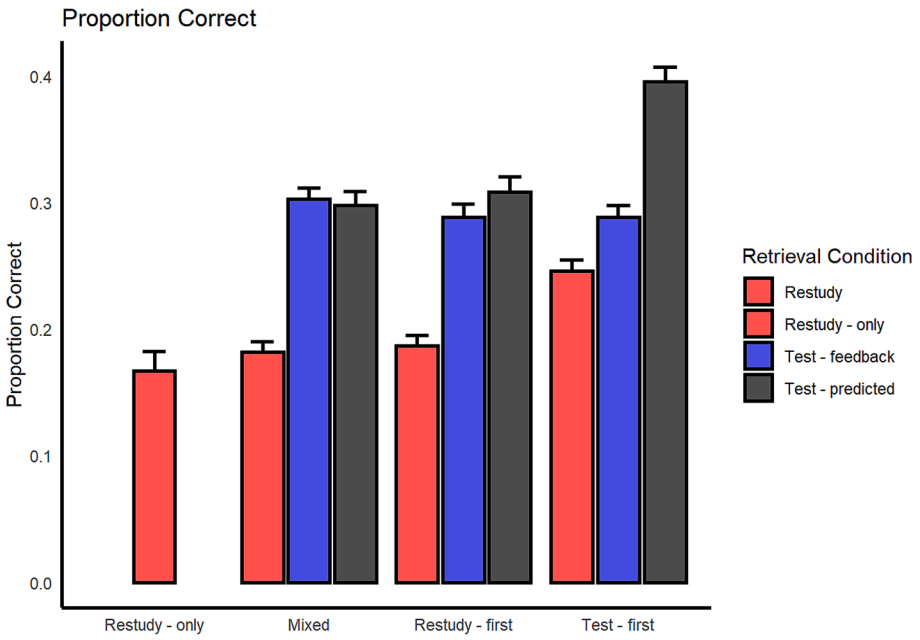

Fig. 5. Mean observed restudy (PC_R) and test (PC_T) condition proportions correct for the Experiment, along with the model prediction for the test conditions, separately for the mixed, restudy-first group, the test-first group. Error bars are standard errors calculated separately for each condition and group.

reflects a larger mean PC_R in the test-first group (.278) than in the restudy-first group (.216), consistent with our FTE confound hypothesis.

Final test model Fits: Means and distributions

For the restudy-first group, there was no significant difference between the mean observed and predicted PC_T values, $t(63) = -0.63$, $p = .53$, $BF_{01} = 5.15$ (see Fig. 3). As expected given the apparent FTE confound, the model fit for the test-first group, based on PC_R for that group, was poor. More critically, however, the model fit for the test-first group based on PC_R for the restudy-first group was good. The observed PC_T for the test-first group was .37 and the predicted-PC_T based on the restudy-first group was .36, $t(134) = 0.27$ $p = .79$, $BF_{01} = 5.2$.

The cumulative distribution results (see Rickard, 2020) for the restudy and test conditions of the two groups are shown in Fig. 4, panels a and b, along with the predicted PC_T distribution based on the restudy condition of each respective group. For the restudy-first group (panel a), the prediction for PC_T is approximately correct across the distribution (mean absolute deviation (*MAD*) over quantiles = .027). It should be noted that, although that distribution fit is somewhat less precise than that for the Pan et al. (2015) data in Fig. 2 (see also Rickard, 2020), that result is expected due to the smaller sample size in the current experiment. For the test-first group (panel b), the distribution fit based on PC_R values from that group is poor (MAD = .075), as observed for the means and as expected in the presence of an FTE confound. However, the cumulative distributions for PC_T in the two groups were similar (Fig. 4, panel c), as expected based on the results for the means. Given the evidence for an FTE confound, the correct model prediction for cumulative PC_T values for both groups is based on the PC_R data from the restudy-first group (shown in both panel a and c). That prediction in fact approximates observed PC_T for both the restudy-first and test-first groups in panel c.

The distribution pattern for the test-first group (panel b of Fig. 4) exhibits a clear asymmetry relative to the pattern for the restudy-first group. Specifically, the TE magnitude for the test-first group is similar to that for both the restudy-first group and the model prediction on the far-left side of the distribution but is markedly smaller around the center and right-side of the distribution. That result appears to be driven solely by the different restudy distribution shapes for the two groups. Given that the cumulative distribution is an expression of individual differences in task ability (along with proportion correct sampling variability; see Rickard, 2020), that result indicates individual differences in the FTE magnitude, a topic to which we return later. As a shorthand going forward, we label the cumulative distribution pattern that was observed for the restudy-first group (and also observed in our prior studies using mixed training) as the *type A* pattern, and the pattern observed for the test-first group as a *type B* pattern.

Experiment 2: Restudy-only, Mixed, Test-first, and restudy-first training comparison

Results from Experiment 1 are consistent with an FTE confound in the test-first group that affected only the restudy condition. Those results leave two questions unaddressed, however. First, although the good model fits both the restudy-first group in Experiment 1 and to mixed groups in our prior experiments suggest that mixed training does not produce an FTE confound, a direct experimental test of that possibility has not been done. Second, the results of Experiment 1 do not provide insight into which of the candidate mechanisms underlies the FTE in the test-first design.

To address those issues, we added two new groups in the current experiment: mixed-training and restudy-only training (in which all 80 items were restudied during training, with no test condition). Based on prior results, we hypothesized good model fits and equivalent performance in the mixed and restudy-first groups, with no evidence for an FTE confound in either. Based on the model assumption that, beyond the FTE confound, learning and performance in the test condition is independent of that in the restudy condition, we hypothesized statistically equivalent PC_T in the restudy-first, mixed, and test-first groups, and as well as equivalent PC_R results in the restudy-only, mixed, and restudy-first groups. The latter prediction is tentative, however, because two accounts of the FTE – the proactive interference and reset of encoding accounts – would most naturally predict monotonically worsening of final test performance from the first to the last trained item during restudy. Because there are 80 items in the restudy-only group and only 40 restudied items in the restudy-first group, those two accounts predict that (1) mean PC_R in the restudy-only group will be smaller than for that in the restudy-first group, and (2) PC_R within the restudy-only group will be smaller for the second 40 trained items than for the first 40 trained items.

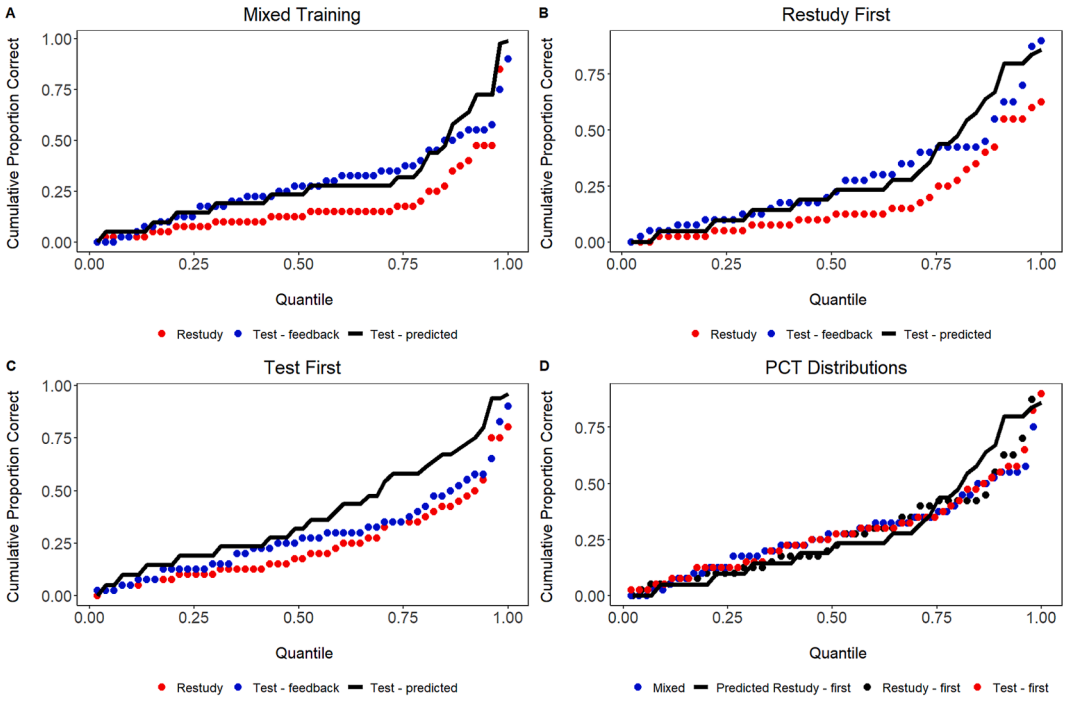

Fig. 6. Cumulative distributions and model fits to the mixed training group (panel A), the restudy first training group (panel B), and the test-first training group (panel C), along with the PC_T results for the mixed and restudy-first groups, along with the predicted PC_T based on the restudy-first group (panel D).

Design, Procedure, and participants

The restudy-first and test-first groups are identical in all respects to those of Experiment 1. The mixed training group is identical to those two groups as well, with the sole exception that the 40 restudied and 40 tested items were randomly intermixed. The restudy-only group is identical to all other groups, with the exception that all 80 items were restudied during the training phase, with no testing condition. Consequently, there was only a restudy condition in that group on the final test. Two hundred and two participants were recruited from the same pool as in Experiment 1 and randomly assigned to the restudy-only group (n = 53), the mixed group (n = 53), the restudy-first group (n = 45), or the test-first group (n = 51).

Results

Training phase means

Mean proportions correct in the tested condition were .22, .2, and .27 in the mixed, restudy-first, and test-first groups, respectively. A between-participants ANOVA indicated no statistically significant differences, $F(2, 146) = 2.59$, $p = .078$.

Final test phase means

The results for mean proportion correction are shown in Fig. 5. Statistical tests on the data (i.e., excluding the model predictions for now) focused on the five predictions of our hypotheses: (1) that the TE will be smaller in the test-first group than in the mixed and restudy first groups, (2) that the TE will not differ for the mixed and restudy-first groups, (3) that PC_R will be larger the test-first group than in the restudy-only, restudy-first, and mixed groups, (4) that PC_R will not differ in the restudy-only, restudy-first, and mixed groups, and (5) that PC_T will not differ in the mixed, restudy-first, and test-first groups.

The first three predictions were evaluated using two-tailed between subjects t-tests. To test the first prediction, we performed a t-test on the TE values (the test minus restudy difference scores), wherein the first group was constituted by the data from the combined restudy-first and mixed groups and the second group was the test-first group: $t(147) = 2.67$, $p = .008$, $d = 0.31$, $BF_{10} = 4.58$. To test the second prediction, we performed an orthogonal t-test comparing the TE magnitude of the mixed group to that of the restudy-first group, with the test-first group excluded: $t(96) = 0.043$, $p = .97$, $d = 0.0061$, $BF_{01} = 4.68$. To test the third prediction, we analyzed the restudy data only, comparing PC_R for the test-first group to that for the combined restudy-only, mixed, and restudy first groups, $t(200) = 2.59$, $p = .01$, $d = 0.26$, $BF_{10} = 3.67$. Those results provide positive evidence consistent with the conclusion that an FTE confound only exists in the test-first group.

We tested the fourth prediction using a one-way between-participants Analysis of Variance (ANOVA) on the PC_R data for the restudy-only, restudy-first, and mixed groups (with the test-first restudy data excluded): $F(2, 148) = 0.178$, $p = .84$, $n^2 = 0.0024$, $BF_{01} = 13.06$. The Bayes factor result indicates positive evidence for no difference in PC_R across those groups. Finally, we performed a between-subjects ANOVA on the PC_T data for the mixed, restudy-first, and test-first groups, again indicating no differences: $F(2, 146) = 0.011$, $p = .99$, $n^2 = 0.00015$, $BF_{01} = 14.91$.

Final test: Model prediction for means

To test the dual-memory predictions for PC_T for the restudy-first and mixed groups, paired-two tailed t-tests were performed between the model prediction for PC_T and the observed PC_T. For neither group did the means for predicted and observed PC_T differ significantly from zero: restudy-first, $t(44) = 0.27$, $p = .79$, $d = 0.028$, $BF_{01} = 6.15$; mixed, $t(52) = 0.39$, $p = .70$, $d = 0.039$, $BF_{01} = 6.59$. To evaluate the model fit to the PC_T data from the test-first group, performed a two-sample t-test, as in Experiment 1, comparing PC_T for the test-first group to the predicted PC_T for the *restudy-first* group: $t(96) = 0.093$, $p = .93$, $d = 0.009$, $BF_{01} = 4.67$.

Final test: Cumulative distribution model predictions

Data and model predictions for the three testing effect groups are shown in Fig. 6. The model fits are approximately correct in the restudy-

first (MAD = .047) and mixed groups (MAD = .044). Note that the sample size is smaller in these groups than for the restudy-first group of Experiment 1, and much smaller than in our prior fits the mixed training data. Given the design equivalence for these current two groups and those prior groups, the poorer distribution fits in Fig. 3 may only reflect higher sampling variability. As expected based on Experiment 1, the distribution fit is poor to the test-first group when based on the restudy data for that group (*MAD* = .101), exhibiting a type B distribution pattern (panel c). Panel D shows the distribution results for PC_T and predicted-PC_T averaged over mixed and restudy-first groups, and the PC_T for the test-first group.

Final Test: Supplementary analysis for the restudy-only group

We evaluated evidence for the proactive interference and reset of encoding accounts of the FTE confound in the test-first group by assessing whether mean PC_R differed for the first 40 and second 40 trained items in the restudy-only group. Mean PC_R was .170 and .152 for those two groups, respectively; the test results were slightly more consistent with the null: $t(52) = 1.72$, $p = .092$, $d = 0.17$, $BF_{01} = 3.33$. Also of note, at the distribution level in the test-first group, the magnitude of the FTE confound was much greater for higher performing participants (as assessed by both PC_R and PC_T) than for lower performing participants, replicating Experiment 1. The proactive interference account would thus predict that the PC_R difference between earlier and later trained items in the restudy-only group should be larger for better performing participants than for worse performing participants (i.e., a larger proactive interference effect for better performing participants). In fact, a median-split of participants in that group based on PC_R yielded negligible differences between those item subsets for both lower performing (.0013) and higher performing (.004) participants. Hence, there appears to be no, or negligible, build-up of proactive interference, nor a reset of encoding effect, across restudy trials.

Discussion

Overall, our hypotheses were supported. There was a robust FTE confound in the test-first group only. The dual-memory model fitted well to PC_T for all three of the groups that involved testing. Further, the test-first group had a distinct Type B cumulative distribution pattern, whereas both the restudy-first and mixed groups had the type A pattern.

Implications for the forward testing effect

The proportion correct mean and distribution patterns observed in Experiments 1 and 2 provide new insight into the FTE. Empirically, those results generalize the FTE confound from free recall (Mulligan et al., 2022) to cued recall. The results also have implications for the theoretical accounts of the FTE and individual differences in the FTE.

Implications for theoretical accounts of the FTE

Our results are inconsistent with both the proactive interference and reset of encoding accounts as applied to cued recall, at least for the stimulus materials used in the current experiment. Both accounts appear to predict an FTE (confound) just as readily in the mixed group as in the test-first group. At the least, the absence of that effect in the mixed group would necessitate an additional mechanism in which the release from proactive interference or reset of encoding following training-phase test trials only occurs when all such trials are blocked prior to restudy. Perhaps more problematic for those accounts is our finding that PC_R in the restudy-only group with 80 items did not differ statistically from that in the restudy-first group with 40 items, and that there was no significant difference on the final test between the first 40 and the second 40 trained items in the restudy-only group.

How, then, can we explain the FTE confound observed in the test-first training design but not the mixed design? Speculatively, we suggest that the FTE for cued recall reflects *enhanced motivation* to learn during restudy that arises only after completing a block of test trials. This account borrows heavily on both the increased effort and strategy change accounts of the FTE. Following a block of test trials, there is an opportunity for participants to reflect on their performance (which is poor by students' academic expectations in most TE studies) allowing for establishment of a new motivational frame to improve learning during subsequent restudy. The resulting FTE may then be the result of increased effort to learn without a shift in strategy (i.e., a higher percentage of the fixed trial time engaged in intentional learning; less mind wandering) or both increased effort and improved strategy selection. In the mixed training design, however, randomly ordered test and restudy trials presented in rapid succession may provide no opportunity for such reflection, yielding no FTE confound. This enhanced motivation account is consistent with results of Soderstrom & Bjork (2014), in which cued-recall testing promoted both more study time and the use of more effective strategies during subsequent restudy.

We cannot determine whether the expectation of a restudy block after the test block in the test-first group caused the FTE confound, because in neither experiment were participants told at the outset that they would engage in alternating periods of studying and testing. The issue of expectation seems pertinent to the mechanisms underlying the FTE and worth exploring in future work. Expectation may facilitate strategic encoding changes, such as using more effective strategies, whereas lack of expectation may imply more automatic mechanisms.

Distribution results and individual differences in the FTE

The type B distribution pattern in the test-first group of both experiments, in which the FTE confound as overall task performance improves, suggests a potent individual difference effect on the FTE magnitude. Before considering that, note that participants with relatively low PC_R and PC_T values have lower episodic ability on average than do participants with high PC_R and PC_T values. That conclusion is supported by correlations observed in multiple studies between an episodic memory assessment and both PC_R and PC_T, yielding correlations ranging from .4 to .6 (Brewer and Unsworth, 2012; Pan et al., 2015; Minear et al.; 2018 Robey, 2019). Hence, the magnitude of the FTE confound in our experiments appears to increase with increased episodic memory ability. Episodic memory ability, however, is a fixed factor that we do not expect to be moderated by having recently performed a difficult memory test, so that factor by itself is unlikely to account for the distinct Type B distribution pattern in the test-first group.

Given that context, there are at least two non-exclusive interpretations of the FTE distribution effect within our speculated enhanced motivation account. First, there may be no, or negligible, individual differences in the degree to which motivation to learn is increased by having performed a test. Rather, the impact of increased motivation during subsequent restudy may be larger for participants with higher episodic memory ability; e.g., increased study time may produce a greater boost in learning for those participants and they may be able use more effective strategies (such as interactive imagery) more effectively. Second, there may be individual differences in motivation to learn following a test that are correlated with episodic memory ability, for reasons that are unknown. We cannot differentiate between those possibilities in the current paper.

Implications for TE individual difference studies

We focus next on the implications of our results for four large-*n* TE datasets in which individual difference factors in the TE were explored, datasets first described in Brewer and Unsworth (2012), Pan et al. (2015), Robey (2019), and Minear et al. (2018). We do not address the TE individual differences findings of those studies directly. Rather, this investigation is (1) pertinent to the capacity those studies to speak

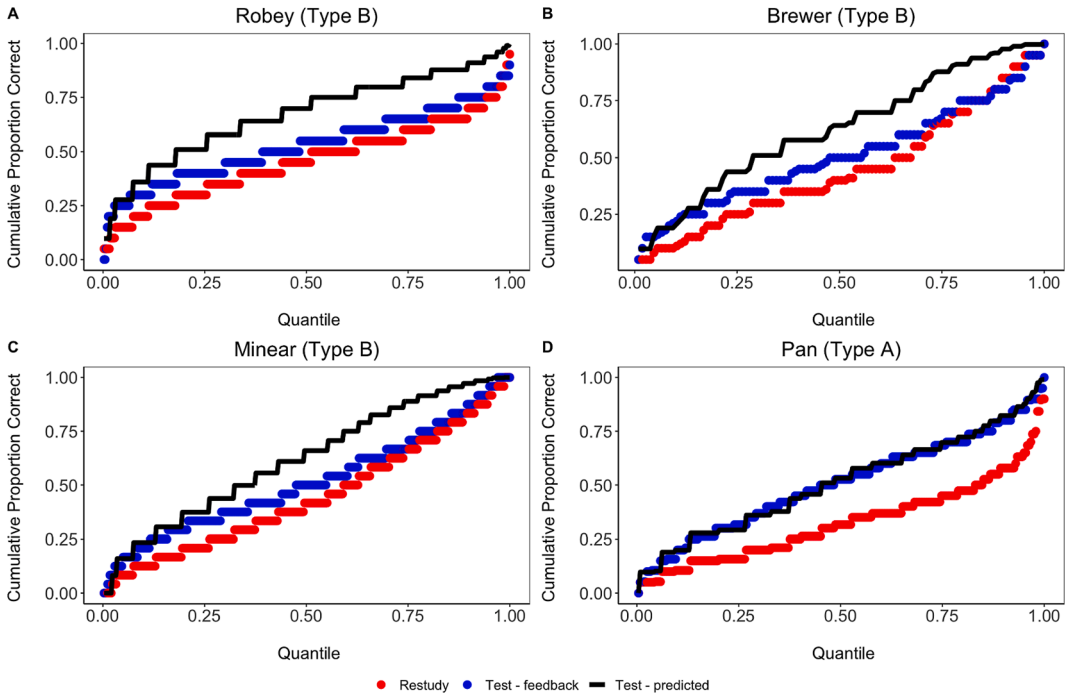

Fig. 7. Final test cumulative distribution model fits to the data from, Robey (2019), Brewer and Unsworth (2012), and Minear et al. (2018), Pan and Rickard (2015); panels a, b, c, and d, respectively. Distribution shape types (Type A or Type B) are indicated at the top of each graph.

unambiguously to TE individual differences, and (2) reinforces the results of the experiments described earlier in the current paper.

All four of those data sets involved the three-phase TE design that we used in Experiments 1 and 2 of the current paper. The major goal of these studies was to gain insight into how the efficacy of testing with feedback might depend on individual difference factors (individual difference factors assessed across those studies were episodic memory ability, working memory ability, attentional control, and intelligence). In all of those studies, the analytical strategy involved calculating the correlation over participants between each individual difference assessment score and the TE. When a statistically significant correlation was observed, it was inferred that the efficiency of testing relative to a restudy control depends on the individual difference factor. That literature has potentially important implications for optimal application of the TE in educational settings.

However, the studies by Brewer and Unsworth (2012), Minear et al., (2018) and Robey (2019) all involved a design in which some form of testing preceded training phase restudy (i.e., they were variants of the current test-first design), raising the prospect, based on the current results, of an FTE confound in their PC_R data. If so, then the observed correlations between individual difference factors and the TE in those studies cannot be interpreted to reflect individual differences in the TE per se, but rather would reflect a correlation between the individual differences and the effect of the combined TE and FTE. In contrast, the Pan et al. (2015) involved mixed training, so that, based on the current results, no FTE confound is expected. We explored the possibility of an FTE confound in those studies by evaluating whether the data yields a Type A or a Type B distribution pattern.

Brewer and Unsworth (2012)

Brewer and Unsworth (2012) explored the cued recall testing effect using English paired-associate materials ($n = 107$). They used a restudy-first blocked training design, so there could be no FTE in the TE experiment proper. However, in the first experimental session their participants completed three individual difference assessments intended to index working memory ability (operation, symmetry, and reading span tasks) prior to the study and training phases of the TE experiment. Although those tasks do not involve episodic memory retrieval, they are error prone tests, and thus are consistent with a more general hypothesis that tests induce an FTE confound during subsequent restudy.

The mean observed PC_R and PC_T values in their TE experiment were .45 and .51, respectively, a much smaller TE than we have observed in cued recall TE experiments wherein there is no FTE (typically around .17). The cumulative distribution results are shown in Fig. 7, panel b, along with the dual-memory model fits for reference. A type B distribution pattern is evident, analogous to that for the test-first groups in Experiment 1 and 2. There thus appears to be an FTE confound in those data.

Comparison of Brewer and Unsworth (2012) and Pan et al. (2015) results

The highly similar experiments of Brewer and Unsworth (2012) and Pan et. al (2015a) constitutes a non-randomized analog to the mixed vs. test-first experimental groups of the current Experiment 2. The Pan et al. (2015a) experiments were designed to mimic closely the Brewer and Unsworth (2012) experiment, in an effort to replicate the findings of those authors regarding episodic memory ability as an individual difference factor in the TE. Those two studies were nearly identical with respect to materials and most aspects of the TE experimental design, with a critical exception being that the Pan et al. study involved mixed training. Thus, we expect that an FTE confound was not present in the Pan et al. study and that performance in the restudy conditions will differ for the two studies, but that performance in test conditions will be roughly equivalent, as in our comparison between mixed and test-first groups in Experiment 2.

That comparison is shown for means in Fig. 8. Despite the large and statistically significant difference in PC_R in the two groups, $t(347) = 4.92$, $p < .0001$, the PC_T values were virtually identical, $t(347) = .05$, $p = .96$, $BF_{01} = 7.8$. Further, the dual-memory prediction for the test condition based on the Pan et al. restudy data fitted well to the mean PC_T data from the Brewer & Unsworth study as well, $t(347) = .13$, $p = .89$, $BF_{01} = 7.7$, again replicating our current experimental finding.

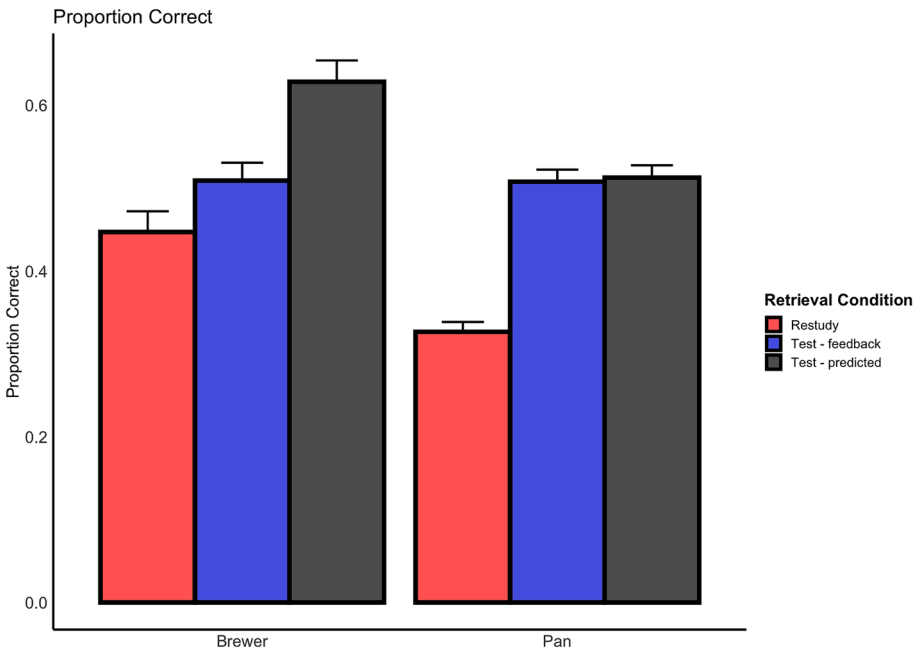

Fig. 8. Mean PC_R, PC_T, and predicted- PC_T values for the Pan et al. (2015) datasets and mean PC_R and PC_T values from the Brewer and Unsworth (2012) datasets.

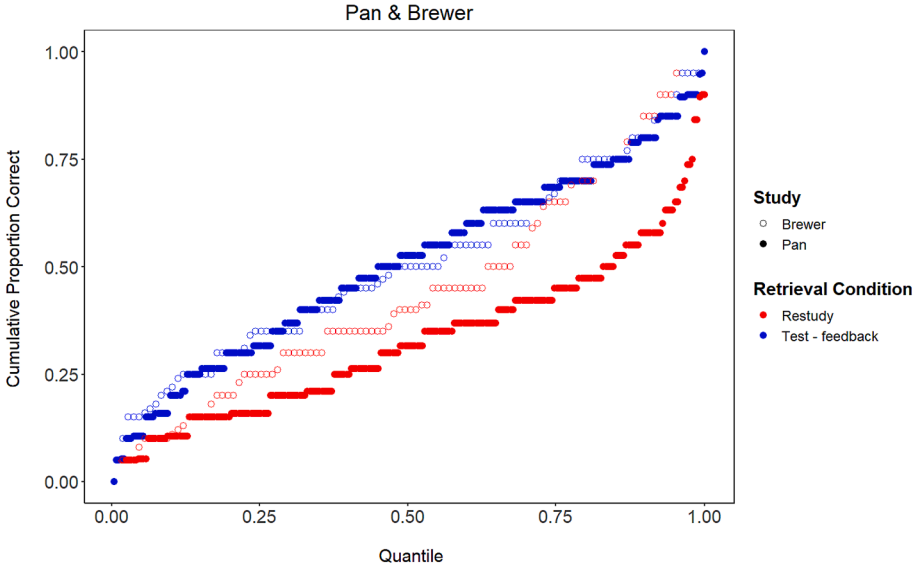

Fig. 9. Final test cumulative distributions for data from both Brewer and Unsworth (2012) and Pan et al. (2015).

Comparison of the cumulative distributions for those two studies (Fig. 9) buttresses the conclusions based on the means. The PC_T distributions in the two experiments are highly similar across quantiles, even though the PC_R distributions diverge markedly. Further, the model predicted PC_T distribution based on the Pan et al. data fitted well to the PC_T distribution of both studies (Fig. 9). It thus appears the groups in the two experiments were similar in their overall task ability. We conclude that the only major difference between groups is the presence of an FTE confound in the Brewer at al. (2012) experiment.

Robey (2019)

Materials in Robey (2019, n = 391) were also English word pairs. Their training design involved, for all pairs, four interleaved blocks of either restudy then test, or test then restudy. Hence, a block of test trials preceded a block restudy trials for all participants, and we expect an FTE confound. The Robey final test results are highly analogous to those for both the Brewer and Unsworth (2012) study and the test-first groups of the current experiments. The mean restudy and test condition proportions correct were .45 and .53, respectively, again indicating a substantially smaller TE than we have observed when there is no FTE confound. Cumulative distribution results, shown in Fig. 7 a, exhibit the type B pattern.

Minear et al. (2018)

Materials in the Minear et al. (2018; $n = 343$) study were 48 Swahili-English word pairs. The training design was similar to that of Robey (2019), involving multiple interleaved blocks of restudy for then test for all pairs, again setting the stage for an FTE confound. The mean PC_R and PC_T values were .45 and .52, respectively. Cumulative distribution results, shown in Fig. 7 c, again exhibit the type B pattern observed in Brewer and Unsworth (2012), Robey (2019), and in the test-first groups of the current Experiments 1 and 2.

Finally, we note a broader potential implication of these results. In the Brewer and Unsworth (2012) paper in particular, it appears that the non-episodic ID assessments themselves caused an FTE confound in the subsequent TE experiment. That finding raises the possibility that, in consecutive administration of two or more ability assessments more generally (i.e., outside of the scope of TE experiments), the earlier administered assessments may yield improved performance on the later administered assessments.

General discussion

We have presented evidence for an FTE confound in the cued-recall testing effect paradigm for the case of test-first blocked training, but not for mixed training or restudy-first training. Further, across the four training groups in Experiment 2 – test-first, restudy-first, mixed, and restudy-only – there were no statistically significant performance differences on the final test for either task other than increased PC_R in the test-first group, with Bayes factors indicating positive support for the null.

The dual-memory model prediction for PC_T held for means, and to an approximation for the distributions, in the restudy-first, mixed, and test-first groups. Those results buttress that model in two respects. First, in three two-group comparisons – the current Experiments 1 and 2 and the Brewer and Unsworth (2012) vs. Pan et al. (2015) datasets – the model fitted well to the PC_T data even when an FTE confound was present in the restudy data, confirmed in each case by applying the model prediction for the group with no FTE confound to the group with an FTE confound. Those findings advance the viability of the model by extending its coverage to cue-recall results in the literature that, in the absence of the FTE confound account, would appear to strongly contradict it. Second, our results confirm the implicit assumption of the model that across the three major training phase designs, there are no, or negligible, task order interactions that affect PC_T.

The results are consistent with our general perspective, in which the mechanisms for the TE and the FTE are entirely independent. Specifically, we propose that the TE reflects independent item-level learning that gives rise to two retrieval routes for tested items (the dual-memory account), whereas the FTE is a task-level phenomenon whereby performing recall enhances motivation to learn during subsequent restudy.

Applied implications

The current results suggest that testing has greater potential to enhance learning than is evident in a subset of experiments in the literature in which some form of testing precedes restudy, including the frequent case in which test and restudy blocks are interleaved and repeated. Among such experiments considered here, the mean TE magnitude was reduced by 40 to 60%, and for the higher performing participants the TE was virtually eliminated. Hence, to fairly assess the practical value of testing for learning, and in related exploration of individual difference factors, TE designs involving mixed or restudy-first training should be preferred.

Conclusions

In summary, the current work makes a number of contributions to research on test-enhanced learning and the forward testing effect. First, we have shown that an FTE confound is present in the cued recall TE paradigm when training phase testing (or testing in prior tasks) is blocked prior to restudy, substantially reducing the TE magnitude. However, we found no evidence for an FTE confound for the case of mixed training. Second, we have shown that the dual memory model predictions generalize to the case in which an FTE confound is present. Third, we have demonstrated that there are systematically different cumulative distribution shapes (type A vs type B) when there is and is not an FTE confound, and that those contrasting distribution patterns indicate potent individual differences in the FTE. Fourth, we have advanced an enhanced motivation account of the FTE for cued recall, which combines prior hypotheses that the FTE is driven by increased effort (Cho et al., 2017) and improved strategy use (Soderstrom & Bjork, 2014; Chan et al., 2018; Cho et al, 2017). Fifth, we have shown that an FTE confound very likely exists in at least three major papers that have explored individual differences in the TE. Finally, we have presented evidence against both the proactive interference and reset of encoding accounts of the FTE for the case of cued recall, at least for the materials employed in the current study.

Footnote

We thank Alison Robey for kindly providing her data and the Brewer and Unsworth (2012) data. We thank Meredith Minear for kindly making their data available.

CRediT authorship contribution statement

Mohan W. Gupta: Conceptualization. **Steven C. Pan:** Conceptualization. **Timothy C. Rickard:** Conceptualization.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

We have made our data, stimuli, and analyses available in an online repository.

Appendix A

Training designs used in the Cued-Recall testing effect literature

To estimate the extent to which a forward testing effect confound may exist in the cued-recall testing effect literature, we examined 56 experiments across 41 papers in that literature that were catalogued in Rickard and Pan (2017). Originally there were 42 papers, but one paper (Metcalfe et al., 2007) was omitted from this analysis due to unique design features that complicated classification. The results revealed that 29% of the experiments implemented testing versus restudy in a *between-participants* design, 34% of the experiments implemented *restudy-first* and *test-first* blocked designs in a counterbalanced manner over participants, ~1.8% (Brewer & Unsworth, 2012) featured a *restudy-first* design, and 36% of the experiments used a *mixed* design with randomly ordered test and restudy trials.

Based on this analysis, 70% of the experiments in the cued-recall testing effect literature were, prior to the experimentation and subsequent analyses conducted in the current study, potentially subject to a forward testing effect confound. Only the experiments that featured a between-participants design were not at risk. For specific results across the analyzed papers and experiments, please refer to Table S1.

Table S1
Training Designs Used in the Cued-Recall Testing Effect Literature.

		Training Phase Arrangement of Testing Versus Restudy			
Reference	**Experiment**	**Between- participants design**	**Counterbalanced *Restudy-first* or *test-first* blocked design**	***Restudy-first* blocked design**	***Mixed* design**
Baghdady et al., 2014	-	X			
Barcroft, 2007	-		X		
Bishara & Jacoby, 2008	Exp 1				X
Brewer & Unsworth, 2012	-			X	
Butler, 2010	Exp 1a		X*		
Carpenter, Pashler, & Vul, 2006	Exp 1				X
	Exp 2		X		
Carpenter et al, 2008	Exp 1				X
	Exp 2				X
	Exp 3				X
Carpenter et al., 2009	-				X
Carrier & Pashler, 1992	Exp 4		X		
Coane, 2013	-	X			
Finley et al., 2011	Exp 2				X
Fritz et al., 2007	Exp 1	X			
	Exp 2	X			
Goossens, Camp, Verkoeijen, & Tabbers, 2014	-		X		
Goossens, Camp, Verkoeijen, Tabbers, & Zwaan, 2014	Exp 1	X			
	Exp 2	X			
Jacoby et al., 2010	Exp 1				X*
	Exp 2				X*
Kang, 2010	Exp 1	X			
	Exp 2		X		
	Exp 3		X		
Kang & Pashler, 2014	Exp 1		X		
	Exp 2		X		
	Exp 3		X		
Kang et al., 2007	Exp 2		X		
Kang et al., 2013	Exp 1		X		
	Exp 2		X		
Karpicke & Blunt, 2011	Exp 1	X			
Keresztes et al., 2014	-		X*		
Kornell & Son, 2009	Exp 1		X**		
Kornell et al., 2011	Exp 2				X
Kromann et al., 2009	-	X			
LaPorte & Voss, 1975	Exp 1	X**			
Lipko-Speed et al., 2014	Exp 1				X
	Exp 2				X
McDermott et al., 2014	Exp 3				X
Morris & Fritz, 2000	Exp 2	X**			
Morris & Fritz, 2002	-	X**			
Morris et al., 2004	Exp 2	X			
Pan, Gopal, et al., 2015	Exp 2				X
	Exp 4				X
Pan, Gopal, & Rickard, 2016	Exp 2				X
	Exp 3				X
Peterson & Mulligan, 2013	Exp 2	X			
Putnam & Roediger, 2013	Exp 2		X**		
Pyc & Rawson, 2010	-	X			
Rohrer et al., 2010	Exp 1		X		
	Exp 2		X		
Rowland & DeLosh, 2015	Exp 4				X
Storm et al., 2014	Exp 2				X
Sumowski et al., 2010	-				X*
Wartenweiler, 2011	Exp 1		X		
Wiklund-Hornqvist et al., 2014	-	X			

Note. (*) indicates that there were some variations in design features beyond that addressed by the four design types listed here. (**) indicates that the design type was inferred from the article text.

Appendix B

The dual-memory framework starts with the simple assumption that learning through exact or near exact study repetition – as on a restudy trial in the TE paradigm – has the effect of reactivating and strengthening the *study memory* that was created during the initial phase. Strengthening is operationalized as increasing the probability correct on a subsequent recall attempt. Hence, study creates a single route to answer retrieval, and restudy strengthens that route.

On each training phase test trial, however, the model assumes that two distinct learning events occur: strengthening of study memory and creation of a separate test memory. Study memory strengthening occurs on both correct test trials and incorrect test trials with feedback. Successful retrieval on the training test must involve reactivation of the study memory that was encoded in the study phase (provided that no pre-experimental associations

that would support that retrieval exist), and that reactivation is assumed to strengthen that study memory, just as restudy does. On incorrect initial test trials, the presented test cue plus the correct answer feedback fully reconstitutes the item as presented during initial study, and thus reactivates and strengthens the study memory, even if the test cue alone was insufficient for that reactivation to occur.

The second learning event on a test with feedback trial is the formation of a separate *test memory*. We assume that, unlike a restudy trial, a test trial is sufficiently distinct from the initial study trial to yield a new and separate memory; on a test trial only the cue is presented, and the presumed task goal is to retrieve the correct response rather than to study for future retrieval. We take those as sufficient conditions for the formation of a separate memory. More specifically, when the cue is presented, an episodic memory is formed that represents the cue in the context of the task set (i.e., *cue memory*). When the correct response becomes available – either through correct retrieval from study memory or via correct answer feedback – an association is created between that cue memory and that response, yielding what we term episodic *test memory*.

That dual routes to retrieval for tested items provides the basis for explaining the testing effect.

The strength of the cue-response association in test memory does not depend on whether the response is correctly retrieved from study memory or provided through feedback; from the "perspective" of the model, only the availability of the correct answer for association with the cue memory matters, not how that availability occurs. Hence, learning on a test trial with feedback is not causally determined by whether the participant's retrieved response is correct or incorrect.

The quantitative model that is based on the framework outlined above was originally designed to apply to the most common cued recall testing effect design, in which there is one study phase trial for each item, and where items in the training phase are randomly assigned to be either restudied or tested with feedback. To create that quantitative model, several simplifying assumptions were made (Rickard & Pan, 2018). At all decision points, the assumption was either supported by prior evidence or, when there was no prior evidence, resulted in the simplest model. Two critical assumptions are that (1) expected *study memory* strength in the experimental design outlined above is the same after training for both restudied items and tested items with feedback, and (2) for tested items, item strengths after training are expected to be the same in *study memory* and *test memory*. However, for each tested item, study and test memory strengths are assumed to be mutually independent (i.e., study memory strength for a particular item after training does not predict test memory strength for that item, and vice versa). Given a few more auxiliary assumptions, Rickard and Pan (2018) showed that, for a hypothetical participant with an infinite number of items randomly assigned to the restudy and test conditions, probability correct in the test condition of the final test (P_T) is given by the inclusive-or equation:

$$P_T = P_{T-s} + P_{T-t} - P_{T-s}P_{T-t} \quad (1)$$

where $P_{T\text{-}s}$ is the probability of correct retrieval through study memory for a randomly selected item in the test condition, and $P_{T\text{-}t}$ is the probability of correct retrieval through the test memory for a randomly selected item in the test condition. As indicated by the foregoing discussion, $P_{T\text{-}s} = P_{T\text{-}t}$ in that equation.

It also follows from the model description above that probability of correct retrieval through study memory in the restudy condition (P_R) is the same as the probability of correct retrieval through study memory in the test condition ($P_{T\text{-}t}$). Hence, in that simplest case quantitative instantiation of the theoretical framework, $P_{T\text{-}s} = P_{T\text{-}t} = P_R$. Equation (1) can thus be reduced to,

$$P_T = P_R + P_R - P_R{*}P_R = 2\,P_R - P_R^2 \quad (2)$$

and the equation for the TE is,

$$P_T = P_R + P_R - P_R{*}P_R = 2\,P_R - P_R^2 \quad (3)$$

The model therefore predicts that both the probability correct in the test condition (P_T) and the TE magnitude are a function solely of probability correct in the restudy condition, P_R. To a close approximation, the predictions of Equations 2 and 3 generalize without modification to proportions correct, as note in the main text.

References

Abel, M., & Roediger, H. L. (2017). Comparing the testing effect under blocked and mixed practice: The mnemonic benefits of retrieval practice are not affected by practice format. *Memory & Cognition, 45*, 81–92.

Agarwal, P. K., Finley, J. R., Rose, N. S., & Roediger, H. L. (2017). Benefits from retrieval practice are greater for students with lower working memory capacity. *Memory, 25*(6), 764–771.

Brewer, G. A., & Unsworth, N. (2012). Individual differences in the effect of retrieval practice from long-term memory. *Journal of Memory and Language, 66*, 407–415.

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.

Carpenter, S. K., Pashler, H., & Vul, E. (2006). What types of learning are enhanced by a cued recall test? *Psychonomic Bulletin and Review, 13*(5), 826–830. https://doi.org/10.3758/BF03194004

Chan, J. C. K., Manley, K. D., Davis, S. D., & Szpunar, K. K. (2018). Testing potentiates new learning across a retention interval and a lag: A strategy change perspective. *Journal of Memory and Language., 102*, 83–96. https://doi.org/10.1016/j.jml.2018.05.007

Tse, C., & Pu, X. (2012). The effectiveness of test-enhanced learning depends on trail test anxiety and working-memory capacity. *Journal of Experimental: Applied, 18*(3), 253–264. https://doi.org/10.1037/a002919

Choi, H., & Lee, H. S. (2020). Knowing is not half the battle: The role of actual test experience in the forward testing effect. *Educational Psychology Review*, dio.org/10.1007/s10648-020-09518-0.

Cho, K. W., Neely, J. H., Crocco, S., & Vitrano, D. (2017). Testing enhances both encoding and retrieval for both tested and untested items. *Quarterly Journal of Experimental Psychology., 2017*(70), 1211–1235. https://doi.org/10.1080/17470218.2016.1175485

Dang, X., Yang, C., & Chen, Y. (2021). Age difference in the forward testing effect : The roles of strategy change and release from proactive interference. *Cognitive Development, 59*(June), Article 101079. https://doi.org/10.1016/j.cogdev.2021.101079

Endres, T., & Renkl, A. (2015). Mechanisms behind the testing effect: An empirical investigation of retrieval practice in meaningful learning. *Frontiers in Psychology, 6*, 1054.

Hopper, W. J., & Huber, D. E. (2018). Learning to recall: Examining recall latencies to test an intra-item learning theory of testing effects. *Journal of Memory and Language, 102*, 1–15.

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

Kornell, N., Bjork, R. A., & Garcia, M. A. (2011). Why tests appear to prevent forgetting: A distribution-based bifurcation model. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 39*(2), 462–472.

McDaniel, M. A., Roediger, H. L., & McDermott, K. B. (2007). Generalizing test-enhanced learning from the laboratory to the classroom. *Psychonomic Bulletin & Review, 14*(2), 200–206.

Minear, M., Coane, J. H., Boland, S. C., Cooney, L. H., & Albat, M. (2018). The benefits of retrieval practice depend on item difficulty and intelligence. *Journal of Experimental Psychology: Learning, Memory, and Cognition., 44*(9), 1474–1486.

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (1998). The University of South Florida word association, rhyme, and word fragment norms. *Available at.* http://www.usf.edu/FreeAssociation/.

Pan, S. C., Gopal, A., & Rickard, T. C. (2016). Testing with feedback yields potent, but piecewise, learning of history and biology facts. *Journal of Educational Psychology, 108*(4), 563–575.

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

Pastötter, B., Schicker, S., Niedernhuber, J., & Bäuml, K. H. T. (2011). Retrieval during learning Facilitates subsequent memory encoding: The forward effect of testing. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 40*, 1039–1048.

Pastötter, B., Tempel, T., & Bäuml, K. H. T. (2017). Long-term memory updating: The reset-of-encoding hypothesis in list-method directed forgetting. *Frontiers in Psychology, 8*(NOV), 1–6. https://doi.org/10.3389/fpsyg.2017.02076

Raftery, A. E. (1995). Bayesian model selection in social research. In P. V. Marsden (Ed.), *Sociological methodology 1995* (pp. 111–196). Blackwell.

Rickard, T. C. (2020). Distribution tests of the dual-memory model of test-enhanced learning. *Psychonomic Bulletin & Review*. in press.

Rickard, T. C., & Pan, S. C. (2018). A dual memory theory of the testing effect. *Psychonomic Bulletin & Review, 25*(3), 847–869.

Rickard, T. C., & Pan, S. C. (2020). Test-enhanced learning for pairs and triplets: When and why does transfer occur. *Memory & Cognition*. https://doi.org/10.3758/s13421-020-01048-y

Robey, A. (2019). The benefits of testing: Individual differences based on student factors. *Journal of Memory and Language, 108*, 1–17.

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.

Schwieren, J., Barenberg, J., & Dutke, S. (2017). The testing effect in the psychology classroom: A meta-analytic perspective. *Psychology Learning & Teaching, 16*(2), 179–196.

Soderstrom, N. C., & Bjork, R. A. (2014). Testing facilitates the regulation of subsequent study time. *Journal of Memory and Language, 73*(1), 99–115. https://doi.org/10.1016/j.jml.2014.03.003

Sohlberg, R., Olsson, F., & Gander, P. (2021). The Effect of Forward Testing as a Function of Test Occasions and Study Material. *Behavioral Science, 11*(114), 1–16.

Szpunar, K. K., McDermott, K. B., & Roediger, H. L. (2008). Testing during study insulates against the buildup of proactive interference. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34*(6), 1392–1399. https://doi.org/10.1037/a0013082

Wagenmakers, E. J., Love, J., Marsman, M., et al. (2018). Bayesian inference for psychology. Part II: Example applications with JASP. *Psychonomic Bulletin & Review, 25*, 58–76.

Weinstein, Y., Gilmore, A. W., Szpunar, K. K., & McDermott, K. B. (2014). The role of test expectancy in the build-up of proactive interference in long-term memory. *Journal or Experimental Psychology: Learning, Memory, & Cognition, 40*(4), 1039–1048. https://psycnet.apa.org/doi/10.1037/a0036164.

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

Yang, C., Potts, R., & Shanks, D. R. (2017). The forward testing effect on self-regulated study time allocation and metamemory monitoring. *Journal of Experimental Psychology: Applied, 23*(3), 263–277. https://doi.org/10.1037/xap0000122

Yang, C., Potts, R., & Shanks, D. R. (2018a). Enhancing learning and retrieval of new information: A review of the forward testing effect. *NPJ science of learning, 3*(1), 1–9.

Yang, C., Zhao, W., Luo, L., Sun, B., Potts, R., & Shanks, D. R. (2021). Testing Potential Mechanisms Underlying Test-Potentiated New Learning. *Journal of Experimental Psychology*. Learning Memory and Cognition, (Advance online publication.

Yang, C., & Shanks, D. R. (2018b). The Forward Testing Effect: Interim Testing Enhances Inductive Learning. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 44*(3), 485–492.