Mem Cogn (2013) 41:429–442
DOI 10.3758/s13421-012-0274-5

Metacognition of the testing effect: Guiding learners to predict the benefits of retrieval

Jonathan G. Tullis · Jason R. Finley · Aaron S. Benjamin

Published online: 14 December 2012

Abstract If the mnemonic benefits of testing are to be widely realized in real-world learning circumstances, people must appreciate the value of testing and choose to utilize testing during self-guided learning. Yet metacognitive judgments do not appear to reflect the enhancement provided by testing Karpicke & Roediger (Science 319:966–968, 2008). In this article, we show that under judicious conditions, learners can indeed reveal an understanding of the beneficial effects of testing, as well as the interaction of that effect with delay (Experiment 1). In that experiment, subjects made judgments of learning (JOLs) for previously studied or previously tested items in either a cue-only or a cue–target context, and either immediately or after a 1-day delay. When subjects made judgments in a cue-only context, their JOLs accurately reflected the effects of testing, both immediately and at a delay. To evaluate the potential of exposure to such conditions for promoting generalized appreciation of testing effects, three further experiments elicited global predictions about restudied and tested items across two study/test cycles (Experiments 2, 3, and 4). The results indicated that learners' global naïve metacognitive beliefs increasingly reflect the beneficial effects of testing when learners experience these benefits with increasing external support. If queried under facilitative circumstances, learners appreciate the mnemonic enhancement that testing provides on both an item-by-item and global basis but generalize that knowledge to future learning only with considerable guidance.

Keywords Testing effect · Metacognition · Monitoring · JOLs · Guided instruction

J. G. Tullis · J. R. Finley · A. S. Benjamin
Department of Psychology,
University of Illinois at Urbana-Champaign,
Urbana, USA

J. G. Tullis (✉)
Department of Psychology,
University of Illinois, 603 E. Daniel St.,
Champaign, IL 61820, USA
e-mail: jtullis2@illinois.edu

Present Address:
J. R. Finley
Department of Psychology,
Washington University in St. Louis,
St. Louis, USA

Guiding learners to predict the benefits of retrieval

For research on learning and memory to be relevant to students who wish to enhance their performance in the classroom, that research must acknowledge the fact that a significant portion of learning occurs outside of the classroom, under the supervision of only the student. In circumstances in which no teacher directly guides the learning activities, learners must rely upon their own metacognition to determine what they need to study, how to study, and when to cease study. Self-regulated aspects of learning have significant implications for the effectiveness of students' learning efforts and achievement in education (Dunlosky & Theide, 1998). For example, how study time is allocated across items often determines how much is remembered (Son & Kornell, 2008; Tullis & Benjamin, 2011a, b). Being an effective learner requires the ability to make appropriate study decisions (e.g., Finley, Tullis, & Benjamin, 2009; Metcalfe, 2009), and the effectiveness of these decisions is directly modulated by the quality of metacognitive monitoring (Metcalfe & Finn, 2008; Thiede, Anderson, & Therriault, 2003). When monitoring judgments are inaccurate or biased, study decisions can result in suboptimal learning (Atkinson, 1972; Kornell & Bjork, 2008; Tullis & Benjamin, 2011a).

In this article, we consider whether learners are sensitive to the mnemonic effects of testing. We will briefly review

the testing effect and consider extant research suggesting that learners fail to accurately monitor the mnemonic effects of testing. After that, we report four experiments that evaluated the extent to which learners *do* accurately monitor the mnemonic effects of testing. Experiment 1 investigated whether learners' metacognitive judgments reflect the mnemonic benefits of testing under conditions that promote judgments based on mnemonic cues rather than naïve theory (Kelley & Jacoby, 1996; Koriat, 1997). Experiments 2, 3, and 4 addressed whether learners attribute improved memory performance to testing and whether this knowledge generalizes to global judgments about future learning.

Metacognition and the testing effect

Retrieval has enormous potential to enhance long-term retention, particularly if learners appreciate its benefits and utilize it properly during self-regulated learning. However, learners' metacognitive judgments fail to reflect the advantages that successful retrieval provides for long-term retention. In order to assess whether learners recognize the benefits that testing provides, researchers have surveyed undergraduate students about their real-life study habits. When students free report the study strategies they use, 11 % report that they practice retrieval, 40 % report using flashcards, and 43 % report practicing solving problems (the latter two options could be viewed as a means of self-testing; Karpicke, Butler, & Roediger, 2009). When choosing the study activities they use from a given list of options, 18 % of students report using self-testing as a means of studying. The percentage of students who report using self-testing when they have an opportunity to restudy afterward increases to 42 %. However, another survey provides a much higher estimate of the use of self-testing by suggesting that up to 71 % of students regularly test themselves with practice problems (Hartwig & Dunlosky, 2012). Even with the highest estimates of self-testing, far from all students report using self-testing to bolster mnemonic performance.

While many learners do not utilize retrieval during study to benefit retrieval, those who do recognize only the indirect benefits of testing. A majority of students who report using self-testing report that they use testing as a means of assessing rather than improving learning, which is an indirect effect of testing. Learners are largely unaware that successful retrieval directly improves learning. When asked "If you test yourself while studying, why do you do it?" approximately two thirds of students report that they test themselves in order to determine what they do and do not know so they can better allocate future study time (Kornell & Bjork, 2007; Kornell & Son, 2009). Only around 20 % of students report that they test themselves because they learn more from testing than from restudying. Researchers have thus argued that learners do not grasp the immense improvement that testing affords for long-term memory retention (Agarwal, Karpicke, Kang, Roediger, & McDermott, 2008; Karpicke, 2009; Karpicke & Roediger, 2008; Kornell & Son, 2009; Roediger & Karpicke, 2006).

Learners' appreciation of the mnemonic effects of retrieval has also been assessed using judgments of learning (JOLs) with controlled stimuli such as word lists and prose passages. Roediger and Karpicke (2006) evaluated metamemory of prose passages by having subjects either study a passage multiple times or study it once and then take tests on the material. Subjects then predicted how well they would remember the passages on a test 1 week later. Subjects rated repeatedly restudied passages as more memorable than tested passages, even though final free recall performance was greater for the tested passages. Kornell and Son (2009) replicated this finding using word pairs in a flashcard-like procedure. Subjects predicted higher levels of recall for word pairs that were restudied versus tested, although final cued recall performance was greater for the tested pairs. In fact, in *all* previous laboratory experiments investigating the effectiveness with which learners monitor the effects of testing, subjects judged restudied items as more likely to be remembered but actually remembered more of the previously tested items. Karpicke and Roediger (2008) thus argued that "students exhibit no awareness of the mnemonic effects of retrieval practice" (p. 968). Consequently, in laboratory paradigms in which learners control their own learning activities, they often choose to restudy rather than test themselves—a counterproductive act that decreases eventual performance on a memory test (Karpicke, 2009). However, in a survey asking about learners' study habits, a large proportion of learners reported using flash cards in order to help memorize information (Wissman, Rawson, & Pyc, 2012). This may reveal a disconnect between control used in artificial laboratory settings, where learners may not value remembering information highly, and that used in classroom settings, where learners may value remembering information more.

Conditions that promote accurate metacognition

Conditions that promote long-term retention through "desirably difficult" processing during acquisition (Bjork, 1999) often yield metacognitive judgments that inappropriately reflect the difficulty of immediate acquisition, rather than the robustness of long-term learning. Learners' failure to appreciate the effects of testing thus parallels similar failures to appreciate the benefits of spacing repetitions (Baddeley & Longman, 1978), interleaved practice (Zechmeister & Shaughnessy, 1980), imagery (Shaughnessy, 1981), and release from proactive interference (Diaz & Benjamin, 2011).

However, in these cases, the opportunity for a *comparative*, *diagnostic* retrieval of material is an important mediating variable in whether learners appreciate the effects of manipulations of learning on long-term memory. Four factors appear to be particularly important. First, JOLs are more sensitive to differences between processing conditions when those conditions are varied within a list than when they are varied between lists or between subjects (Begg, Duft, Lalonde, Melnick, & Sanvito, 1989; Shaw & Craik, 1989). Presumably, such conditions promote metacognitive accuracy because they invite direct comparison between the learning conditions, a comparison that is more difficult between lists (and even more difficult between subjects!). Second, the type of metacognitive monitoring question posed influences the accuracy of the resultant judgments. Aggregate JOLs may depend upon different, often less diagnostic, cues than do item-by-item JOLs (Mazzoni & Nelson, 1995). Third, adding a delay between study and JOLs improves the predictive accuracy of JOLs (Dunlosky & Nelson, 1994; Nelson & Dunlosky, 1991) because retrieval at a delay is more diagnostic of later retrieval probability than is immediate recall (Benjamin & Bjork, 1996). Finally, the information present when subjects make JOLs greatly impacts JOL accuracy: JOLs made when both the cue and target from a word pair are present are significantly less accurate at predicting future cued recall than are predictions made when only the cue is present (Dunlosky & Nelson, 1992, 1994). It is likely that inclusion of the target during the JOL solicitation either dissuades subjects from attempting to retrieve the target or renders such retrieval nondiagnostic (Benjamin, Bjork, & Schwartz, 1998); consequently, when the target is present, subjects fall back on analytic, naïve theories about memorability, rather than the more accurate mnemonic strategy otherwise used (Jacoby & Kelley, 1987).

When judgments are made in contrastive, delayed, item-by-item, cue-only contexts, many failures of metacognitive monitoring are allayed. For example, subjects accurately predict the mnemonic benefits of interactive imagery, spacing, and the number of presentations (Begg et al., 1989; Carroll, Nelson, & Kirwan, 1997; Dunlosky & Nelson, 1994). The effects of testing, like most other desirably difficult learning conditions, may be metacognitively misunderstood unless care is taken to elicit judgments under contrastive, delayed, cue-only conditions. Experiment 1 compared metacognitive monitoring of the testing effect under facilitative conditions as outlined above with that under the more adverse conditions that are normally used to collect these judgments

Experiment 1

Extant data all seem to show a serious failure of metacognitive monitoring of the testing effect. However, these experiments all have characteristics that make the metacognitive demand on the learner quite high. First, all previous studies investigating metacognitive monitoring of the testing effect have elicited JOLs immediately after the final study/test session. This is an important detail, not only because delaying judgments enhances their accuracy, but also because at such short test intervals, restudying actually *is* beneficial, when compared with testing (Roediger & Karpicke, 2006). Thus, these JOLs may *accurately* reflect the current consequences of testing, which are negative, rather than the future mnemonic strength of items when they will be tested. We elicited JOLs both shortly after testing and at a longer delay in order to determine whether learners can recognize both the mnemonic costs and benefits that testing provides. Second, all previous studies have manipulated testing between subjects or via blocked-list designs. In this study, previously tested and restudied items were varied within a list so as to enable comparative judgments. Third, previous testing effect studies have relied upon aggregate JOLs, asking subjects to predict final memory performance across groups of tested and groups of restudied items. Prior studies gathered aggregate JOLs instead of item-by-item JOLs in order to avoid confounding learning conditions (restudy vs. testing) with item-by-item JOL context (cue-only vs. cue–target). However, by only gathering aggregate JOLs, prior studies have encouraged subjects to base their judgments on analytic naïve theories about learning conditions, rather than on each item's subjective mnemonic cues (Koriat, 1997). We solicited JOLs on an item-by-item basis in a phase following the restudying/testing manipulation so that subjects could rely upon subjective processing cues for each item when making JOLs. Finally, we elicited JOLs in both cue-only and cue–target contexts in order to demonstrate that a metacognitive appreciation of the effects of testing depends on the opportunity to engage in diagnostic retrieval of the queried material. We predicted that learners would appreciate the mnemonic costs and benefits of testing when JOLs were elicited in facilitative, delayed, cue-only conditions. Our predictions of metacognitive ratings were less clear in cue–target conditions, where learners were reexposed to all the word pairs after the crucial testing/restudy manipulation had occurred. The added exposure to both the restudied and tested word pairs might unpredictably alter both memory predictions and final performance.

Method

Subjects

One hundred twenty introductory-level psychology students from the University of Illinois at Urbana-Champaign participated in exchange for partial course credit.

Materials

Materials were 64 unassociated English word pairs (e.g., *viking–napkin*). All cues and targets were nouns and ranged in length from four to eight letters. Target words were selected for high concreteness (M = 572, SD = 32) and high imageability (M = 578, SD = 34), using the Medical Research Council Psycholinguistic Database (Coltheart, 1981). Cues varied in written frequency (Kučera & Francis, 1967) from 0 to 286 (M = 43.5, SD = 73.8). Targets varied in written frequency from 1 to 591 (M = 120.8, SD = 138.5). Lack of association between cues and targets was confirmed using the University of South Florida Word Association, Rhyme, and Word Fragment Norms (Nelson, McEvoy, & Schreiber, 1998).

Design

The experiment used a 2 × 2 × 2 mixed design, in which practice condition was manipulated within subjects (phase 2 = restudy vs. test), JOL context was manipulated between subjects (phase 3 = cue only vs. cue–target), and retention interval was manipulated between subjects (delay between phases 2 and 3 = none vs. 1 day). A schematic of the four between-subjects groups is displayed in Table 1. For each condition, n = 30.

Procedure

Subjects were run individually on computers programmed with MATLAB using the Psychophysics Toolbox extensions (Brainard, 1997). The procedure consisted of four phases: initial presentation, practice, JOLs, and final test. Subjects were alternately assigned to the four conditions on the basis of the order in which they were run.

Phase 1: initial presentation Subjects were shown 32 word pairs randomly chosen from the pool of 64 pairs. Pairs were presented one at a time for 4 s each, with an interstimulus interval of 0.5 s. Presentation order was determined randomly for each subject.

Phase 2: practice Subjects restudied half (16) of the word pairs and were tested on the other half. Word pairs were randomly assigned to practice condition for each subject, with the constraint that exactly half of the word pairs were restudied and half were tested. On restudy trials, the word pairs were simply re-presented for 4 s each, followed by a 0.5-s interstimulus interval. On test trials, the cue (left-hand) word of a pair was presented, and subjects were instructed to type the corresponding target (right-hand) word or to type a question mark if they could not remember the target word. There was no time limit for responding, and each trial was followed by a 0.5-s interstimulus interval. Restudy trials and test trials were interleaved, with order determined randomly for each subject.

Delay Half of the subjects continued the experiment with no delay between phases 2 and 3. The other half of the subjects left the lab after phase 2 and returned the following day to complete phases 3 and 4.

Phase 3: JOLs Subjects were randomly assigned to make individual JOLs for all 32 word pairs in either a cue–target context or a cue-only context. The JOL phase was separated from the restudy/test phase by either no delay or a long (1-day) delay. Trial order was determined randomly for each subject.

For subjects in the cue–target context, on each trial they were shown the entire word pair for 4 s and then given the following JOL prompt: "How sure are you that you will remember this item on the upcoming final test?" Subjects responded using a scale ranging from 1 (*I am sure I will NOT remember this item*) to 4 (*I am sure I WILL remember this item*). The presented word pair remained visible during the judgment. There was no time limit for responding, and each trial was followed by a 0.5-s interstimulus interval.

Table 1 The four between-group conditions in Experiment 1

	Day 1				Day 2	
Condition	Phase 1	Phase 2	Phase 3	Phase 4	Phase 3	Phase 4
Cue-only JOL, no delay	Study	½ restudy ½ test	Cue-only JOLs	Test		
Cue–target JOL, no delay	Study	½ restudy ½ test	Cue–target JOLs	Test		
Cue-only JOL, 1-day delay	Study	½ restudy ½ test			Cue-only JOLs	Test
Cue–target JOL, 1-day delay	Study	½ restudy ½ test			Cue–target JOLs	Test

For subjects in the cue-only context, on each trial they were shown the cue (left-hand) word of a pair and were instructed to type the corresponding target (right-hand) word or to type a question mark if they could not remember the target word. There was no time limit for responding. Once they had responded, subjects were given the same JOL prompt as that used in the cue–target context, with the same response scale. The presented cue word and the subject's response remained visible during the judgment. There was no time limit for responding, and each trial was followed by a 0.5-s interstimulus interval.

Phase 4: final test Subjects were given cued recall test trials on all 32 word pairs, using the same procedure as the test trials in the practice phase. Order was determined randomly for each subject.

Results

First, we will consider the results for subjects who made JOLs in the presence of the cue only. Then we will consider results for subjects who made JOLs in the presence of the cue–target pair. All statistics reported here are significant at an $\alpha < .05$ level unless otherwise noted. Effect sizes for comparisons of means are reported as Cohen's *d* calculated using the pooled standard deviation of the groups being compared (Olejnik & Algina, 2000, Box 1 Option B). Effect sizes for ANOVAs are reported as $\hat{\omega}^2_{partial}$ calculated using the formulae provided by Maxwell and Delaney (2004).

Cue-only JOL context

Memory performance In this and all subsequent experiments, subjects' answers were considered correct only if the answer matched the target exactly. The mean proportion of items recalled during the phase 2 practice test was .49 (*SD* = .26). Figure 1 shows mean final cued recall performance and mean JOLs for subjects in the cue-only JOL context. A 2 (mode of practice) × 2 (delay) mixed model ANOVA showed that cued recall performance significantly declined across the 1-day delay, $F(1, 58) = 24.99$, $p < .001$, $\omega^2_{partial} = .424$. Critically, mode of practice interacted with delay, $F(1, 58) = 43.96$, $p < .001$, $\omega^2_{partial} = .086$, such that restudied items were recalled at a higher rate than tested items on the immediate test, $t(29) = 4.12$, $p < .001$, $d = 0.60$, and tested items were recalled at a higher rate than restudied items after the 1-day delay, $t(29) = 5.75$, $p < .001$, $d = 0.66$.

Judgments of learning A 2 (mode of practice) × 2 (delay) mixed model ANOVA was performed on JOLs. The JOLs, shown in the bottom panel of Fig. 1, clearly reflect the pattern apparent in actual memory performance: They show both the decline of performance across the delay, $F(1, 58) = 10.58$, $p < .001$, $\omega^2_{partial} = .372$, and the interaction between mode of practice and delay, $F(1, 58) = 27.60$, $p < .001$, $\omega^2_{partial} = .046$, such that restudied items were given higher JOLs than were tested items at no delay, $t(29) = 2.73$, $p = .011$, $d = 0.36$, and tested items were given higher JOLs than were restudied items at the 1-day delay, $t(29) = 4.90$, $p < .001$, $d = 0.34$.

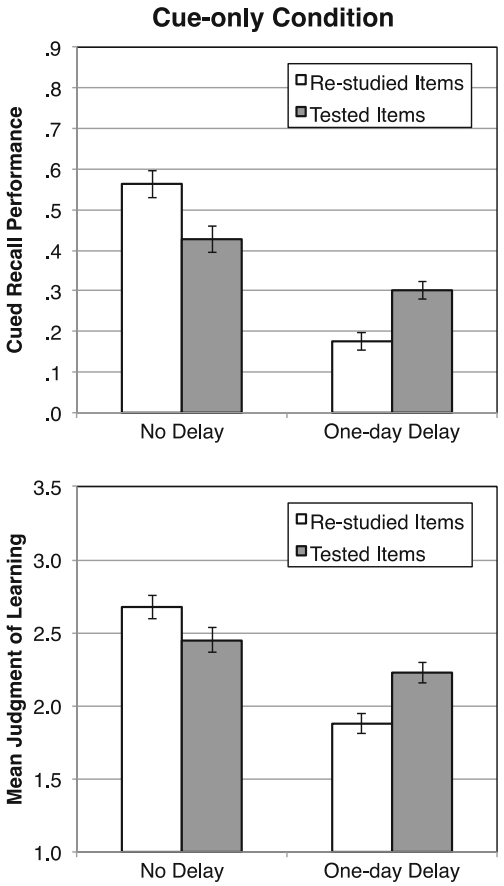

Fig. 1 Mean proportion correct recall on the final cued recall test (top panel) and mean judgment of learning on a scale of 1 to 4 (bottom panel) for cue-only subjects as a function of retention interval and mode of practice during phase 2 (Experiment 1). Error bars represent the standard error of the difference scores within each delay condition

Cue–target JOL context

Memory performance Figure 2 shows mean final cued recall performance and mean JOLs for subjects in the cue–target JOL context. Performance did not change across mode of practice, $F(1, 58) = 1.89$, $p = .17$, $\omega^2_{partial} = .002$, or delay, $F(1, 58) = 0.04$, $p = .84$, $\omega^2_{partial} < 0$, and the interaction between delay and mode of practice was not significant, $F(1, 58) = 3.24$, $p = .08$, $\omega^2_{partial} = .005$. The added restudy opportunity during phase 3 washed out the testing effect from

the prior study phase, particularly when the restudy (phase 3) and test (phase 4) events took place the second day, since there was no delay between the last re-presentation of the list and the final test.

Judgments of learning As is shown in Fig. 2, JOLs given during restudy in phase 3 were slightly but significantly higher for tested items, $F(1, 58) = 21.00$, $p < .001$, $\omega^2_{partial} = .031$, but the JOLs did not differ across the retention intervals, $F(1, 58) = 0.001$, $p = .97$, $\omega^2_{partial} < 0$. They do not reflect the interactive pattern evident in Fig. 2, $F(1, 58) = 1.91$, $p = .17$, $\omega^2_{partial} = .001$, nor should they, since the conditions do not support that interaction in actual retention.

Discussion

Under conditions that allowed for diagnostic retrieval from long-term memory immediately prior to judgments, JOLs accurately reflected the testing effect across retention intervals. Subjects in the cue-only condition accurately accorded restudied items higher JOLs immediately after practice but accorded tested items higher JOLs at a delay. Thus, subjects showed sensitivity to the mnemonic strength imparted by testing and the interaction of that effect with retention interval, thereby implicitly recognizing the differential rates of forgetting following study and test trials.

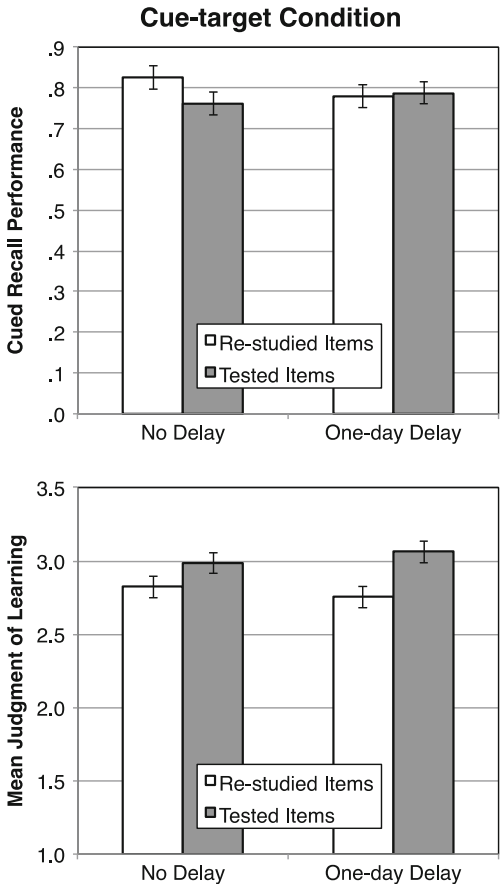

Fig. 2 Mean proportion correct recall on the final cued recall test (top panel) and mean judgment of learning on a scale of 1 to 4 (bottom panel) for cue–target subjects as a function of retention interval and mode of practice during phase 2 (Experiment 1). Error bars represent the standard error of the difference scores within each delay condition

Although judgments mirrored the effects of testing when they were elicited in a cue-only context, they did not when elicited in the presence of both the target and the cue. In a cue–target context, subjects cannot (or do not) infer the effectiveness of learning from the outcome of a concurrent retrieval attempt; rather, they are limited to less diagnostic cues, such as their memories for the outcomes of prior tests (Finn & Metcalfe, 2008). To assess whether subjects relied upon the outcome of the prior phase 2 test to determine their phase 3 JOLs, we computed gamma correlations between phase 2 test performance and phase 3 JOLs within each subject. This correlation was high in both the cue-only ($G = .84$) and the cue–target ($G = .86$) conditions, indicating either a reliance on the outcome of the prior test for the later JOL or, more simply, an item effect whereby well-learned items were more likely to have been recalled during phase 2 and were accorded higher JOLs in phase 3. The interesting results concern the relationship between phase 3 JOLs and final test performance, which was much higher for the cue-only ($G = .93$) than for the cue–target ($G = .54$) condition, $t(110) = 5.13$. This result suggests that the high correlations between phase 2 performance and phase 3 JOLs may have had a different basis for the two prediction contexts. Under cue-only conditions, subjects relied on the outcome of the phase 3 retrieval attempt to make JOLs, which was highly diagnostic of later performance. Under cue–target conditions, they relied upon the outcome of a prior test (one made during phase 2) to make JOLs, which was not as accurate of a basis.

Relying upon outcomes of prior tests to make cue–target JOLs, however, results in better metacognitive resolution than when no prior tests have been attempted. Learners exhibit very accurate metacognitive resolution in cue-only conditions because they base JOLs on the outcomes of concurrent diagnostic retrieval attempts (tests of each item; Dunlosky & Nelson, 1994). This experiment shows that an added benefit of testing is an improvement in resolution of subsequent metacognitive judgments made in cue–target contexts. The metacognitive resolution of cue–target JOLs in predicting final memory performance, as measured by signal detection theoretic value d_a (Benjamin & Diaz, 2008), was significantly higher for previously tested than previously restudied items both at no delay (1.10 vs. 0.64), $t(29) = 2.31$, and delay (1.08 vs. 0.68), $t(29) = 2.08$. This extends the *memory for past test heuristic* literature (Finn &

Metcalfe, 2008) by indicating that reliance upon the outcomes of prior tests when JOLs are made is a wise strategy: Metacognitive resolution is improved in cue–target contexts if learners can base metacognitive judgments on their memory for prior test outcomes. Furthermore, this reveals that students' reliance upon tests to assess what they know and do not know (Kornell & Son, 2009) is a reasonable study strategy. Basing judgments on the results of prior tests allows learners to improve their future metacognitive accuracy, which may benefit learners' study choices.

Previously tested word pairs elicited higher JOLs than did previously studied words in the cue–target condition even though final recall did not differ, which, although not central to our claims here, differs from extant results in the literature. The cue–target condition utilized within this experiment, however, significantly differs from those in the present literature. While the prior testing effect literature has usually relied upon aggregate JOLs as indicators of metacognitive monitoring, we solicited item-by-item JOLs for both studied and tested items within each subject. Since the judgments were elicited within each subject on an item-by-item basis, the cue–target condition reexposed learners to all the word pairs following the restudy/test manipulation. Learners who were not able to retrieve items during the phase 2 test were provided an additional opportunity to learn the correct target for the word pair during the phase 3 cue–target JOLs. The additional exposure to the word pairs likely altered both predictions and performance across conditions. Furthermore, item-by-item monitoring may have allowed subjects to base metacognitive judgments upon subjective cues, like cue familiarity, rather than upon naïve analytic theories about the influence of testing. The reliance on subjective rather than analytic cues may have caused subjects to provide higher predictions for previously tested items than for studied items, because cue familiarity, an influential subjective cue (Metcalfe, Schwartz, & Joaquim, 1993), for tested items may have been greater than that for restudied items. This is likely because subjects spent significantly more time responding to each cue during a test trial during phase 2 than they spent viewing an item during a representation trial (6.2 vs. 4 s), $t(119) = 3.95$. The extra time spent processing the cue may have augmented cue familiarity, resulting in higher metacognitive predictions without affecting future memory performance (Metcalfe et al., 1993).

This result adds to and qualifies the existing research on metacognition of the testing effect. Extant research shows that analytic, theory-based judgments about the effects of testing are, at best, incomplete and likely inaccurate, and these results show that mnemonic-based judgments of the testing effect are very accurate. Under favorable circumstances, metacognitive judgments accurately reflect the mnemonic costs and benefits of testing, as seen in the first experiment. Learners can accurately judge the mnemonic strength imparted by testing. However, it remains unclear whether learners attribute mnemonic differences in performance to prior restudy/test conditions and whether they will generalize this knowledge to future testing predictions. In the next three experiments, we assessed learners' global metacognitive judgments about tested and restudied items to determine whether, and under what circumstances, learners would attribute improved memory performance to prior testing. It is of considerable practical interest to know whether subjects can use the accurate mnemonic-based judgments to credit testing with improved memory performance and, consequently, generalize this knowledge to predict mnemonic benefits of testing during future study episodes. Furthermore, the circumstances under which learners translate accurate item-by-item judgments into accurate aggregate judgments remain unknown. JOLs can reflect a variety of different cues and not necessarily reveal whether learners attribute mnemonic performance to differing study activities (Koriat, 1997). Aggregate judgments may more closely reflect knowledge about study activities than do item-by-item JOLs.

In these experiments, we examined changes in learners' metacognitive beliefs about the effectiveness of testing by measuring learners' global predictions of the mnemonic consequences of testing and restudying across a pair of study/test cycles. Even though learners have significant misconceptions about the mnemonic consequences of many aspects of learning, they often modify those beliefs to correctly reflect the differential effectiveness of strategies used to encode items (Brigham & Pressley, 1988), the differential effectiveness of cues used to retrieve items (Bieman-Copland & Charness, 1994), and how item characteristics affect memorability (Benjamin, 2003; Tullis & Benjamin, 2011b) through task experience.

Experiment 2

In Experiments 2, 3, and 4, we collected global predictions about performance on restudied and tested items to assess general, theory-based beliefs about the differential mnemonic effectiveness of restudying, as compared with testing. We collected these global predictions across two study/test cycles and measured changes in predictions after direct experience with the task introduced in Experiment 1. If learners successfully update their predictions concerning the effectiveness of these practice activities, they should make higher predictions for tested than for restudied items during the second cycle. In the following three experiments, learners predicted how many restudied and tested items they would remember the following day. On the second day, learners were tested on all previously practiced items, practiced a second list of items, and made predictions about how many restudied and tested items from the second list they would remember the following day. Our

focus was on how the global predictions about the effectiveness of restudy and testing would change after experience with the task. Performance on list 2 was not measured, because we were interested solely in the circumstances that prompt learners to predict benefits of testing.

We varied the metacognitive burden placed on learners across Experiments 2, 3, and 4. In Experiment 2, a high metacognitive burden was placed on learners: In order to use knowledge from the first list to change predictions on the second, learners had to remember the study condition for each item at a long delay and had to track their memory performance between the study conditions. In Experiment 3, we reduced the metacognitive burden learners face by providing partial external support. Learners were told about the study conditions for each item during the list 1 test. Finally, in Experiment 4, we reduced the metacognitive burden even more by additionally providing an aggregate summary of performance comparing tested and restudied items. The high metacognitive burden in Experiment 2 might prevent learners from effectively updating their knowledge to predict benefits of testing, while the reduced metacognitive burden in Experiments 3 and 4 might facilitate this change.

Method

Subjects

Thirty-five introductory-level psychology students from the University of Illinois at Urbana-Champaign participated in exchange for partial course credit.

Materials

The same word pair pool was used as in Experiment 1. The word pair pool was randomly split into two equal halves, such that one half (32 word pairs) was studied and practiced on day 1 and the other half was studied and practiced on day 2.

Design

The experiment used a 2 × 2 within-subjects design. The independent variables were practice condition (phase 2 = restudy vs. test), and list number (1 vs. 2). Global predictions of memory performance for restudied items and for tested items were collected following practice on each list.

Procedure

The procedure was similar to that used in Experiment 1 but included an additional cycle of study, practice, and predictions for a new word list on day 2. Additionally, Experiment 2 collected global metamnemonic predictions instead of item-by-item JOLs in order to assess general beliefs about the effectiveness of each practice technique.

The initial study and practice phases in Experiment 2 were identical to phases 1 and 2 in Experiment 1, but the prediction phase differed. Following the practice phase on day 1, all subjects made global predictions about how many restudied items and how many tested items they would remember the next day. Subjects were asked, "From the 16 word pairs that you RE-STUDIED, how many do you think you will remember tomorrow?" and "From the 16 word pairs that you were TESTED on, how many do you think you will remember tomorrow?" The order of the questions alternated between consecutive subjects. Subjects returned a day later to complete the cued recall test.

Once they had finished the cued recall task for the first list, subjects completed another cycle of presentation and practice phases with a new list of 32 word pairs. Finally, as they did after practicing the first list, subjects made global predictions about how many restudied and how many tested items they thought they would remember from the second list on the following day. No second test was administered.

Results

Memory performance

The mean proportion of items recalled during the practice test phase of the first list was .36 (*SD* = .23). On the final test for the first list of items, subjects recalled a greater proportion of tested items (*M* = .24, *SD* = .17) than of restudied items *(M* = .19, *SD* = .15), $t(34) = 2.31$, $p = .027$, $d = 0.31$. Of the 35 subjects, 19 showed a testing effect, 8 showed no differences between restudied and tested items, and 8 showed an advantage for restudied over tested items. The mean proportion of items recalled during the practice test phase of the second list was .41 (*SD* = .24).

Global predictions

The mean global predictions are shown in Fig. 3 (top panel). A 2 (prediction cycle: first or second) × 2 (study activity: restudy or test) ANOVA was conducted on the global predictions, and no significant main effects were found. Furthermore, a significant interaction was not found between prediction cycle (first or second list) and study activity (restudy or test), $F(1, 34) = 0.44$, $p = .51$, $\omega^2_{partial} < 0$. When we restricted the data to only those subjects who showed a testing effect, the same pattern of results held: A significant interaction between prediction cycle and study activity was not found, $F(1, 18) = 0.00$, $p = .99$, $\omega^2_{partial} = 0$.

Discussion

Subjects did not update their knowledge to reflect the testing effect through direct experience with the task. This finding is in contrast to prior studies that examined updating of the effects of item characteristics (Benjamin, 2003; Tullis & Benjamin, 2011b) or encoding regimens (Dunlosky & Hertzog, 2000). However, there are results revealing failures to update with experience (Diaz & Benjamin, 2011), and there are parallels to be drawn between that study and this one. In both tasks, a heavy burden was placed on subjects to accurately remember, at test, the origin of individual items—in this case, which items had been previously studied and which had been tested. The long delay between practice and test in the present experiment makes that task even more difficult. In addition to remembering which practice method was used for each item, learners must accurately tally the number of items correctly remembered for each type of practice utilized. By this explanation, the major bottleneck to subjects learning about the testing effect comes from the difficulty of tracking the relationship between the conditions of learning and test performance.

Alternatively, subjects may have based their global predictions on the same considerations as their item-by-item JOLs. That is, they may not seek to assess the relationship between study conditions and test because they already have a theory about the effects of testing (an incorrect one). If so, subjects predict greater recall for restudied items than for tested items for the same reasons that restudying leads to higher immediate cue-only JOLs than testing typically does.

In Experiment 3, we eliminated some of the extraneous cognitive load in order to evaluate these competing explanations. We reduced the cognitive demands of the knowledge updating task by introducing partial feedback for subjects; we told subjects which type of study activity was used for each item and whether they were correct on the final test of the first list. If learners predicted benefits of testing during the second list in Experiment 3, it would indicate that the difficulty of tracking the relationship between prior study activities and eventual test performance was likely to be the cause of the failed knowledge updating in Experiment 2, and not mere persistence with an incorrect and unmalleable mental model about the effects of testing on memory.

Experiment 3

In Experiment 3, we introduced partial feedback to subjects by providing them with information about their final answers' correctness and about which study activity was used for each item.

Method

Subjects

Fifty-three introductory-level psychology students from the University of Illinois at Urbana-Champaign participated in exchange for partial course credit.

Materials and design

The materials and design were identical to those used in Experiment 2.

Procedure

The procedure was very similar to that used in Experiment 2, but the test included more information about each subject's performance and the study activity used during the practice phase. Feedback given during the cued recall test on the second day included whether the subject's response was correct or incorrect and whether a given item had been restudied or tested during the practice phase. Feedback remained on the screen for 3 s before it was removed and the next item was tested.

Results

Memory performance

The mean proportion of items recalled during the initial practice test phase was .42 (SD = .24). Final cued recall performance was higher for previously tested items (M = .21, SD = .18) than for previously restudied items (M = .15, SD = .17), $t(52) = 3.14$, $p = .003$, $d = 0.34$. Of the 53 subjects, 35 showed a testing effect, 7 showed no differences between restudying and testing, and 11 showed a mnemonic advantage of restudy over test. The mean proportion of items recalled during the practice phase of the second list was .46 (SD = .27).

Global predictions

Subjects' predictions of performance are displayed in Fig. 3 (middle panel). A 2 (prediction phase: first or second) × 2 (study activity: restudy or test) ANOVA on global predictions shows a marginal interaction between prediction phase and study activity, $F(1, 52) = 3.23$, $p = .08$, $\omega^2_{partial} = .002$, and no main effects. Subjects rated restudying as numerically more effective than testing during the first prediction cycle, $t(52) = 1.12$, $p = .27$, $d = 0.09$, but rated testing as numerically more effective than restudying during the second, $t(52) = 1.47$, $p = .15$, $d = 0.12$.

An analysis of the subset of 35 subjects who showed a testing effect reveals the same pattern of data. The 2

(prediction phase) × 2 (study activity) ANOVA on predictions failed to show a significant interaction between phase and study activity, $F(1, 34) = 1.68$, $p = .20$, $\omega^2_{partial} < 0$. Subjects rated restudying as numerically better during the first list and testing better during the second, but neither of these post hoc tests reached significance, $t(34) = 0.55$, $p = .59$, $d = 0.06$, and $t(34) = 1.56$, $p = .13$, $d = 0.07$, respectively.

Discussion

Subjects may have slightly updated their knowledge to reflect the testing effect through direct experience with the task and experimenter-provided feedback. In contrast to Experiment 2, the ordering of predictions favored restudying in the first list but favored testing in the second. The size of knowledge updating may have been very small and variable across these lists, so the interaction between cycle and study activity reached only marginal significance. We still placed a large cognitive burden on learners, since they had to tally the number of correctly answered restudied and tested items. Providing information about whether each item was restudied or tested may provide more impetus to change metacognitive ratings than does feedback about a response's correctness. Other studies have shown that providing information at test about prior study condition by blocking similar items helps update metacognitive knowledge (Price, Hertzog, & Dunlosky, 2008), presumably because learners can more easily count how many items from each condition were recalled. Tallying the correctly answered items from each condition may be the last impediment to knowledge updating, since learners may lack the cognitive resources to do so on their own. In Experiment 4, we eliminated this extraneous cognitive load by providing all of the feedback of Experiment 3 and a tally of the number of restudied and tested items they successfully recalled. If learners predicted benefits of testing during the second list in Experiment 4, it would indicate that the difficulty of tallying the number of items correctly recalled prevented knowledge updating about the effectiveness of self-testing.

Experiment 4

In Experiment 4, we provided even more external support by tallying the number of correct restudied and tested items during the first list before subjects started the second.

Method

Subjects

Twenty-eight introductory-level psychology students from the University of Illinois at Urbana-Champaign participated in exchange for partial course credit.

Materials and design

The materials and design were identical to those used in Experiments 2 and 3.

Procedure

The procedure was very similar to that used in Experiment 3, but the test included more information about each subject's performance and the study activity used during the practice phase. Once all items were tested, subjects were given global feedback about their performance that indicated how many restudied items out of 16 they had correctly answered and how many tested items out of 16 they had correctly answered. The order of this global differentiated feedback was randomized and remained on the screen until the subjects indicated that they were ready to proceed.

Results

Memory performance

The mean proportion of items recalled during the initial practice test phase was .36 (SD = .24). Final cued recall performance was higher for previously tested items (M = .27, SD = .21) than for previously restudied items (M = .17, SD = .16), $t(27) = 3.45$, $p = .002$, $d = 0.54$. Of the 25 subjects, 19 showed a testing effect, 6 showed no differences between restudy and test, and 3 showed an advantage of restudy over test. The mean proportion of items recalled during the practice phase of the second list was .55 (SD = .23).

Global predictions

Subjects' predictions of performance are displayed in Fig. 3 (bottom panel). A 2 (prediction phase: first or second) × 2 (study activity: restudy or test) ANOVA on global predictions shows a significant interaction between prediction phase and study activity, $F(1, 27) = 11.01$, $p = .003$, $\omega^2_{partial} = .017$, and no main effects. Subjects rated restudying as more effective than testing during the first prediction cycle, $t(27) = 3.48$, $p = .071$, $d = 0.27$, but rated testing as marginally more effective than restudying during the second, $t(27) = 1.67$, $p = .099$, $d = 0.25$. An analysis of the subset of 19 subjects who showed a testing effect revealed the same interaction between prediction phase and study cycle, $F(1, 17) = 9.57$, $p = .006$, $\omega^2_{partial} = .024$, as the larger group of subjects. Additionally, post hoc tests reveal that this subset of subjects initially predicted that restudying would lead to better performance than would testing, $t(18) = 2.28$, $p = .035$, $d = 0.24$, but during the second cycle, reversed their ratings to predict that

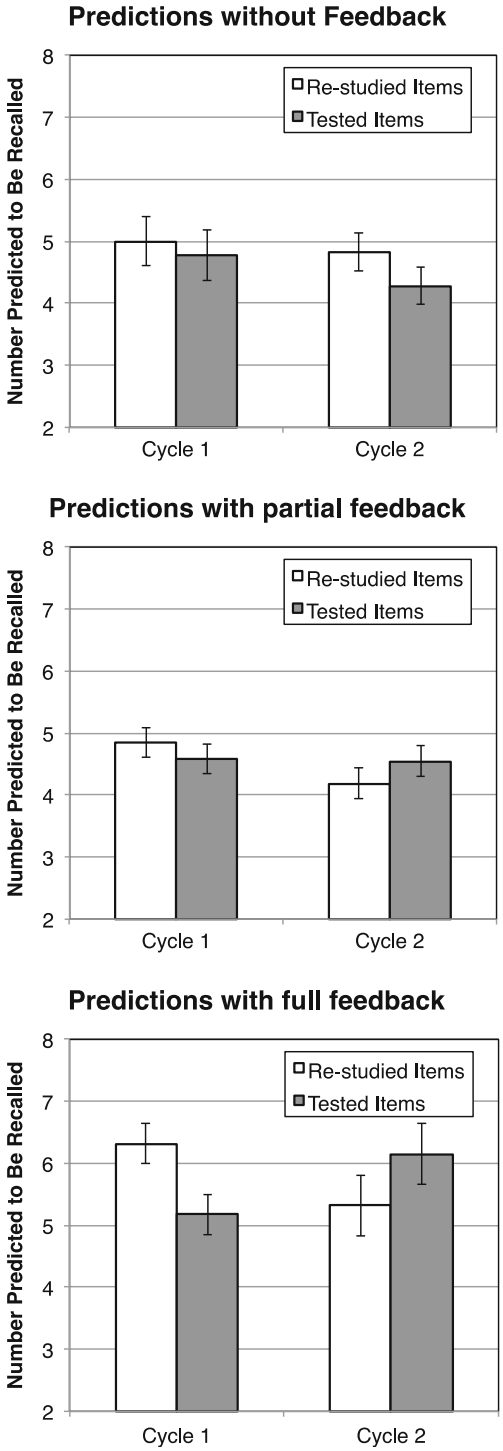

Fig. 3 Number of items predicted to be recalled (out of 16 possible) as a function of study/test cycle and study activity. Top panel shows judgments provided with no external support (Experiment 2); middle panel shows judgments with some external support (Experiment 3); bottom panel shows judgments provided with extensive external support (Experiment 4). Error bars represent the standard error of the difference scores within each cycle

testing would produce better memory performance than would restudying, $t(18) = 2.43$, $p = .026$, $d = 0.41$.

Discussion

Subjects significantly updated their beliefs about the advantages of testing over restudying when facilitated by external support. Initially, subjects predicted that restudying would lead to superior memory than would testing; through task experience and with feedback, subjects rated testing as greater for promoting long-term memory. This shift suggests that the failure to accurately update knowledge in Experiments 2 and 3 was due to a failure to accurately track the effects of testing and restudying on items tested after a 24-h delay and tally the results accordingly.

Results across Experiments 2 through 4

We combined the results across Experiments 2, 3, and 4 to see how varying the amount of external support influences learners' predictions across cycles. To do so, we classified learners' predictions on lists 1 and 2 as to whether they predicted an advantage for testing, equivalent performance, or an advantage for restudying. We determined whether this classification matched with their actual mnemonic performance on list 1 across the three feedback conditions. As is shown in Fig. 4, the proportion of learners whose predictions matched their list 1 performance increased across the cycles only in the full feedback condition. Learners' predictions seem to have been greatly influenced by the tally of their performance and less influenced by the partial feedback given in Experiment 3.

Next, we used learners' metacognitive classifications (predicting a testing effect, equivalent performance, or a restudy advantage) across lists 1 and 2 to categorize learners' predictions as shifting toward or away from the testing effect. For example, if learners predicted equivalent performance on the first cycle but predicted a testing effect on their second cycle, their

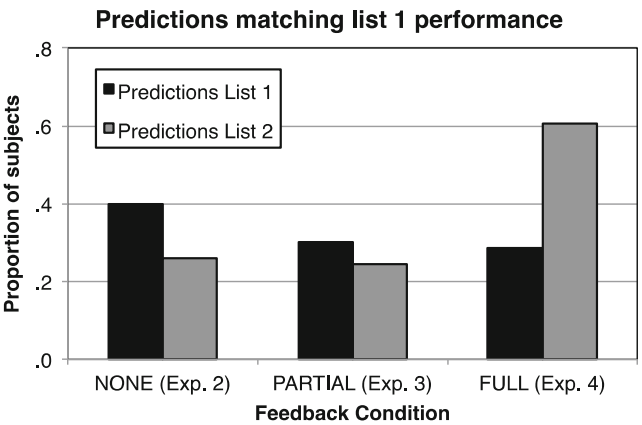

Fig. 4 The proportion of learners whose predictions on lists 1 and 2 match their mnemonic performance on list 1 by feedback condition (Experiment 2 = none; Experiment 3 = partial; Experiment 4 = full)

predictions were classified as shifting toward the testing effect. Conversely, if learners predicted equivalent performance on the first cycle but a restudy study advantage on the second, their predictions were classified as shifting away from the testing effect. The proportions of learners shifting toward and away from the testing effect are displayed in Fig. 5. With increasing external support, the proportion of learners shifting toward the testing effect increases, while the proportion of learners shifting away from the testing effect drops.

The capability of learners to recognize the benefits of differing study activities may depend upon the magnitude of those benefits. Across Experiments 2, 3, and 4, learners showed a relatively small testing effect, such that tested items were remembered significantly, but only slightly, better than restudied items. Differences in final mnemonic performance between prior tested and restudied items may have been small because successful retrieval on the initial test was somewhat lower than in other studies (approx. 40 %). Low levels of initial successful retrieval reduce the overall advantage for the class of tested items. Increasing successful retrieval during the initial test by increasing the amount of prior learning may increase the size of the testing effect and change what benefits of testing learners recognize. Learners may have struggled to detect the advantages of testing because the mnemonic advantage of testing was small throughout these experiments. Consequently, learners need supportive conditions to value testing more than restudying. If a larger testing effect exists, learners may show knowledge updating under less supportive conditions.

General discussion

The results here demonstrate that when subjects make judgments under conditions that promote diagnostic self-testing and provide differentiated feedback, their JOLs show sensitivity to the delay-contingent memory effects of testing versus restudying. Experiment 1 shows that the accuracy of metacognitive judgments for tested items is very similar to metacognitive accuracy for other beneficial study habits, like imaging items and spacing study between repetitions. The accuracy of metacognitive judgments in all cases depends greatly upon the conditions under which the judgments are solicited.

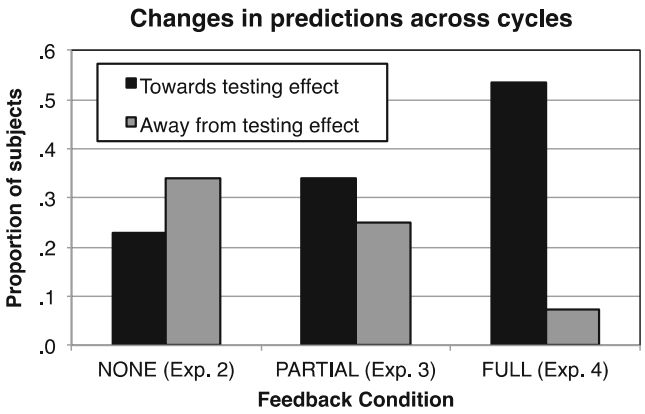

Fig. 5 The proportion of learners whose predictions shifted toward or away from valuing testing more than restudying across the two list cycles by feedback condition (Experiment 2 = none; Experiment 3 = partial; Experiment 4 = full)

The ability to accurately appreciate the benefits provided by study activities carries vital consequences for learners. Accurate metacognition drives effective study behavior and can improve cognition without significant increases in work or effort, while inaccurate or biased metacognition may lead learners to adopt suboptimal study behaviors and lead to inferior levels of learning. Learners' metacognitions are very accurate under circumstances in which mnemonic assessment is feasible, contrastive, and diagnostic prior to metacognitive judgments.

However, taking full advantage of such conditions is difficult, in part, because the benefits of testing do not arise until long after study choices have been made. In order to benefit from control over learning, learners must recognize the advantages of testing when making choices about study activities. Experiments 2, 3, and 4 investigated the conditions under which such knowledge updating can occur. In Experiment 2, a lack of change in global memory predictions about the effectiveness of testing, as compared with restudying, across study lists revealed that learners do not update beliefs about the effects of testing when they bear the full burden of tracking the relationship between items and their learning conditions. However, as the experiments provided more and more external support to track these relationships, global memory predictions shifted to predict beneficial effects of retrieval. With adequate support, learners can learn through experience to predict the long-term advantages that testing provides. Other studies also suggest that students recognize the benefits that testing provides when they are provided with extensive external support about study conditions during testing (Einstein et al. 2012). Furthermore, after students recognize the benefits of testing, they report that they incorporate more testing into their self-controlled study activities (Einstein et al., 2012).

The results described in these experiments might help provide a bridge between the many results indicating failures of metacognitive monitoring with respect to the testing effect and suggestive evidence that metacognitive control may reflect the advantages of testing. Learners who experience the mnemonic benefits of generating words (a cognitive activity similar but not identical to testing; Karpicke & Zaromb, 2010) enhance their future processing of read items to render those items as well remembered as generated items (deWinstanley & Bjork, 2004). In deWinstanley and Bjork's study, learners were presented with short textbook passages

that contained both to-be-generated and to-be-read critical items. Learners' memory for critical items was measured by fill-in-the-blank tests across two study/test cycles. Learners' memory performance showed a generation advantage during the first cycle but lacked a generation advantage during the second. The absence of the generation advantage during the second cycle occurred because memory performance for the to-be-read items improved to the level of the to-be-generated items. The authors argued that subjects recognized the mnemonic advantage of generation during the first test and spontaneously developed processing strategies for to-be-read material during the second cycle that rendered to-be-read material as well remembered as to-be-generated material. DeWinstanley and Bjork's results suggest that when learners recognize the benefits that *testing* provides, they may spontaneously engage in better encoding and processing strategies for restudied items during subsequent lists in order to boost memory performance for those items. Subsequent processing and performance changes may be apparent on restudied items, where learners could ignore (or actively obscure) the presence of the target in order to test themselves and improve their memory performance. However, we cannot evaluate this claim in our data, since memory performance on the second list was not measured.

This set of experiments provides insight into what is required for learners to update their metacognitive strategy knowledge. First, learners' monitoring must be accurate enough to notice a consistent difference in acquisition or performance between learning conditions. Larger differences in performance may draw more attention to the differences in conditions and allow learners to notice the differences. Next, learners must attribute performance differences to the study conditions and must not disregard differences as arising from idiosyncratic item characteristics. Attributing differences in performance to study conditions requires tracking the study conditions of individual items and tallying the performance of those items across each condition. The ability to track and tally performance, while still recalling items during the test, may require significant working memory resources. The support that the environment provides may modulate the amount of working memory resources needed to track and tally the items accurately. Finally, to modify existing beliefs about strategies, learners must acknowledge the discrepancy between their prior beliefs and actual results.

Interventions designed to educate learners about the effect of testing need to consider the tracking and tallying burden that hampers learning about testing effects and provide support in such a manner as to allay that burden. The extent to which conditions support such tracking and tallying will determine whether attempts to enlighten metacognition will founder or succeed. The extensive feedback detailing performance on tested, as compared with restudied, items does not naturally occur in the course of classroom learning because teachers cannot segregate performance as a function of the study activities chosen by the learners during self-regulated learning. The need to fully instruct and support learners concerning the effectiveness of metacognitive strategies mirrors current conceptions about fully guiding learners about content knowledge (Klahr & Nigam, 2004). Many argue that learners do not efficiently or effectively learn content knowledge when required to discover or construct it on their own (Sweller et al. 2011). The present study shows that learners may not effectively learn about the differential effectiveness of metacognitive strategies unless provided with full, explicit guidance. This combination of circumstances suggests that automated tutors (see Finley et al., 2009) may be the best source of hope for providing the detailed feedback necessary to translate the appreciation of testing evident in Experiment 1 into the successful theory-based predictions evident in Experiment 4.

Author note This research was funded in part by Grant R01 AG026263 from the National Institutes of Health.

References

Agarwal, P. K., Karpicke, J. D., Kang, S. H. K., Roediger, H. L., & McDermott, K. B. (2008). Examining the testing effect with open- and closed-book tests. *Applied Cognitive Psychology, 22,* 861–876.

Atkinson, R. C. (1972). Optimizing the learning of a second-language vocabulary. *Journal of Experimental Psychology, 96,* 124–129.

Baddeley, A. D., & Longman, D. J. A. (1978). The influence of length and frequency of training session on the rate of learning to type. *Ergonomics, 21,* 627–635.

Begg, I., Duft, S., Lalonde, P., Melnick, R., & Sanvito, J. (1989). Memory predictions are based on ease of processing. *Journal of Memory and Language, 28,* 610–632.

Benjamin, A. S. (2003). Predicting and postdicting the effects of word frequency on memory. *Memory & Cognition, 31,* 297–305.

Benjamin, A. S., & Bjork, R. A. (1996). Retrieval fluency as a metacognitive index. In L. Reder (Ed.), *Implicit Memory and Metacognition*. Mahwah: Erlbaum.

Benjamin, A. S., Bjork, R. A., & Schwartz, B. L. (1998). The mismeasure of memory: When retrieval fluency is misleading as a metamnemonic index. *Journal of Experimental Psychology: General, 127,* 55–68.

Benjamin, A. S., & Diaz, M. (2008). Measurement of relative metamnemonic accuracy. In J. Dunlosky & R. A. Bjork (Eds.), *Handbook of Memory and Metamemory* (pp. 73–94). New York: Psychology Press.

Bieman-Copland, S., & Charness, N. (1994). Memory knowledge and memory monitoring in adulthood. *Psychology & Aging, 9,* 287–302.

Bjork, R. A. (1999). Assessing our own competence: Heuristics and illusions. In D. Gopher & A. Koriat (Eds.), *Attention and performance XVII. Cognitive regulation of performance: Interaction of theory and application* (pp. 435–459). Cambridge: MIT Press.

Brainard, D. H. (1997). The psychophysics toolbox. *Spatial vision, 10* (4), 433–436.

Brigham, M. C., & Pressley, M. (1988). Cognitive monitoring and strategy choice in younger and older adults. *Psychology & Aging, 3,* 249–257.

Carroll, M., Nelson, T. O., & Kirwan, A. (1997). Tradeoff of semantic relatedness and degree of overlearning: Differential effects on metamemory and on long-term retention. *Acta Pscyhologica, 95,* 239–253.

Coltheart, M. (1981). The MRC psycholinguistic database. *The Quarterly Journal of Experimental Psychology A: Human Experimental Psychology, 33*(4), 497–505.

deWinstanley, P. A., & Bjork, E. L. (2004). Processing strategies and the generation effect: Implications for making a better reader. *Memory & cognition, 32*(6), 945–955.

Diaz, M., & Benjamin, A. S. (2011). Effects of proactive interference (PI) and release from PI on judgments of learning. *Memory & Cognition, 39,* 196–203.

Dunlosky, J., & Nelson, T. O. (1992). Importance of the kind of cue for judgments of learning (JOL) and the delayed-JOL effect. *Memory & Cognition, 20,* 374–380.

Dunlosky, J., & Nelson, T. O. (1994). Does the sensitivity of judgments of learning (JOLs) to the effects of various study activities depend upon when the JOLs occur? *Journal of Memory and Language, 33,* 545–565.

Dunlosky, J., & Hertzog, C. (2000). Updating knowledge about encoding strategies: A componential analysis of learning about strategy effectiveness from task experience. *Psychology and Aging, 15,* 462–474.

Dunlosky, J., & Theide, K. W. (1998). What makes people study more? An evaluation of factors that affect self-paced study. *Acta Psychologica, 98,* 37–56.

Einstein, G. O., Mullet, H. G., & Harrison, T. L. (2012). The testing effect: Illustrating a fundamental concept and changing study strategies. *Teaching of Psychology, 39,* 190–193

Finley, J. R., Tullis, J. G., & Benjamin, A. S. (2009). Metacognitive control of learning and remembering. In M. S. Khine & I. M. Saleh (Eds.), *New Science of Learning: Cognition, Computers and Collaboration in Education*. New York: Springer Science & Business Media.

Finn, B., & Metcalfe, J. (2008). Judgments of learning are influenced by memory for past test. *Journal of Memory and Language, 58,* 19–34.

Hartwig, M. K., & Dunlosky, J. (2012). Study strategies of college students: Are self-testing and scheduling related to achievement? *Psychonomic Bulletin & Review, 19,* 126–134.

Jacoby, L. L., & Kelley, C. M. (1987). Unconscious influence of memory for a prior event. *Personality and Social Psychology Bulletin, 13,* 314–336.

Karpicke, J. D. (2009). Metacognitive control and strategy selection: Deciding to practice retrieval during learning. *Journal of Experimental Psychology: General, 138,* 469–486.

Karpicke, J. D., Butler, A. C., & Roediger, H. L. (2009). Metacognitive strategies in student learning: Do students practice retrieval when they study on their own? *Memory, 17,* 471–479.

Karpicke, J. D., & Roediger, H. L. (2008). The critical importance of retrieval for learning. *Science, 319,* 966–968.

Karpicke, J. D., & Zaromb, F. M. (2010). Retrieval mode distinguishes the testing effect from the generation effect. *Journal of Memory and Language, 62,* 227–239

Kelley, C. M., & Jacoby, L. L. (1996). Adult egocentrism: Subjective experience versus analytic bases for judgment. *Journal of Memory and Language, 35,* 157–175.

Klahr, D., & Nigam, M. (2004). The equivalence of learning paths in early science instruction: Effects of direct instruction and discovery learning. *Psychological Science, 15,* 661–667

Koriat, A. (1997). Monitoring one's own knowledge during study: A cue utilization approach to judgments of learning. *Journal of Experimental Psychology: General, 126,* 549–570.

Kornell, N., & Bjork, R. A. (2007). The promise and perils of self-regulated study. *Psychonomic Bulletin & Review, 14,* 219–224.

Kornell, N., & Bjork, R. A. (2008). Optimizing self-regulated study: The benefits and costs of dropping flashcards. *Memory, 16,* 125–136.

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

Kucera, H., & Francs, W. N. (1967). *Computational analysis of present-day American English*. Providence: Brown University Press.

Maxwell, S. E., & Delaney, H. D. (2004). *Designing experiments and analyzing data: A model comparison perspective* (2nd ed.). Mahwah: Lawrence Erlbaum Associates.

Mazzoni, G., & Nelson, T. O. (1995). Judgments of learning are affected by the kind of encoding in ways that cannot be attributed to the level of recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21,* 1263–1274.

Metcalfe, J. (2009). Metacognitive judgments and control of study. *Current Directions in Psychological Science, 18,* 159–163.

Metcalfe, J., & Finn, B. (2008). Evidence that judgments of learning are causally related to study choice. *Psychonomic Bulletin & Review, 15,* 174–179.

Metcalfe, J., Schwartz, B. L., & Joaquim, S. G. (1993). The cue-familiarity heuristic in metacognition. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 19,* 851–861.

Nelson, T. O., & Dunlosky, J. (1991). When people's judgments of learning are extremely accurate at predicting subsequent recall: The "delayed-JOL effect. *Psychological Science, 2,* 267–270.

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

Olejnik, S., & Algina, J. (2000). Measures of effect size for comparative studies: Applications, interpretations, and limitations. *Contemporary Educational Psychology, 25,* 241–286.

Price, J., Hertzog, C., & Dunlosky, J. (2008). Age-related differences in strategy knowledge updating: Blocked testing produces greater improvements in metacognitive accuracy for younger than older adults. *Aging, Neuropsychology, and Cognition, 15,* 601–626.

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

Schaughnessy, J. J. (1981). Memory monitoring accuracy and modification of rehearsal strategies. *Journal of Verbal Learning and Verbal Behavior, 20,* 216–230.

Shaw, R. J., & Craik, F. I. M. (1989). Age differences in predictions and performance on a cued recall task. *Psychology and Aging, 4,* 131–135.

Son, L. K., & Kornell, N. (2008). Research on the allocation of study time: Studies from 1890 to the present (and beyond). In J. Dunlosky & R. A. Bjork (Eds.), *A Handbook of Memory and Metamemory* (pp. 333–351). Hillsdale: Psychology Press.

Sweller, J., Ayres, P., & Kalyuga, S. (2011). Cognitive Load Theory. New York: Springer

Thiede, K. W., Anderson, M. C. M., & Therriault, D. (2003a). Accuracy of metacognitive monitoring affects learning of texts. *Journal of Educational Psychology, 95,* 66–73.

Tullis, J. G., & Benjamin, A. S. (2011a). On the effectiveness of self-paced learning. *Journal of Memory and Language, 64,* 109–118.

Tullis, J. G., & Benjamin, A. S. (2011). The effectiveness of updating metacognitive knowledge in the elderly: Evidence from metamnemonic judgments of word frequency. Psychology and Aging. Advance online publication. doi:10.1037/a0025838

Wissman, K. T., Rawson, K. A., & Pyc, M. A. (2012). How and when do students use flashcards? *Memory, 20,* 568–579.

Zechmeister, E. B., & Shaughnessy, J. J. (1980). When you know that you know and when you think that you know but you don't. *Bulletin of the Psychonomic Society, 15,* 41–44.