Psychon Bull Rev (2013) 20:1239–1245
DOI 10.3758/s13423-013-0434-z

BRIEF REPORT

Repeated retrieval practice and item difficulty: Does criterion learning eliminate item difficulty effects?

Kalif E. Vaughn · Katherine A. Rawson · Mary A. Pyc

Published online: 16 April 2013

Abstract A wealth of previous research has established that retrieval practice promotes memory, particularly when retrieval is successful. Although successful retrieval promotes memory, it remains unclear whether successful retrieval promotes memory equally well for items of varying difficulty. Will easy items still outperform difficult items on a final test if all items have been correctly recalled equal numbers of times during practice? In two experiments, normatively difficult and easy Lithuanian–English word pairs were learned via test–restudy practice until each item had been correctly recalled a preassigned number of times (from 1 to 11 correct recalls). Despite equating the numbers of successful recalls during practice, performance on a delayed final cued-recall test was lower for difficult than for easy items. Experiment 2 was designed to diagnose whether the disadvantage for difficult items was due to deficits in cue memory, target memory, and/or associative memory. The results revealed a disadvantage for the difficult versus the easy items only on the associative recognition test, with no differences on cue recognition, and even an advantage on target recognition. Although successful retrieval enhanced memory for both difficult and easy items, equating retrieval success during practice did not eliminate normative item difficulty differences.

K. E. Vaughn · K. A. Rawson
Kent State University, Kent, OH, USA

M. A. Pyc
Washington University, St. Louis, MO, USA

K. E. Vaughn (✉)
Department of Psychology, Kent State University,
P.O. Box 5190, Kent, OH 44242-0001, USA
e-mail: kvaughn4@kent.edu

Keywords Human memory

Substantial research has established that retrieval practice benefits memory, particularly when retrieval attempts are successful (see Rawson & Dunlosky, 2011). Although successful retrieval promotes memory, it remains unclear whether successful retrieval promotes memory equally well for items of varying difficulty. Will easy items outperform difficult items on a final test if both are correctly recalled equal numbers of times during practice?

In prior research manipulating item difficulty, retention has been poorer for difficult than for easy items (e.g., Cull & Zechmeister, 1994). However, most experiments investigating item difficulty utilized fixed numbers of practice trials (e.g., three per item). With a fixed number of trials, difficult items are correctly recalled less often during practice than are easy items. For example, Bahrick and Hall (2005) presented learners with easy and difficult Swahili–English word pairs for five test–restudy trials. On the last trial, performance was lower for difficult than for easy items (67 % vs. 84 %). Not surprisingly, the final test performance one week later was significantly lower for difficult than for easy items (30 % vs. 52 %). Thus, a fixed number of trials favors easy items (due to more successful retrievals during practice), making it difficult to ascertain whether the poorer retention for difficult than for easy items is due to item difficulty or to differential levels of recall success during practice.

Rather than practicing for a fixed number of trials, students typically self-test until items are correctly recalled at least once (Wissman, Rawson, & Pyc, 2012). Similarly, our design ensured that both difficult and easy items were correctly recalled the same number of times during practice. Given that successful recall enhances retention, a logical

prediction would be that equating the numbers of correct recalls during practice would yield similar final test performance for difficult and easy items.

Of course, equating the numbers of successful recalls during practice does not necessarily result in equivalent amounts of learning between difficult and easy items. In line with this possibility, some theoretical accounts have stated that not all successful retrievals are equally beneficial. For example, the *retrieval effort hypothesis* (REH; Pyc & Rawson, 2009) states that on trials in which an item is correctly recalled, successful retrieval enhances memory to a greater extent when the retrieval is more rather than less effortful (to foreshadow a point that will be important later, the REH is a purely descriptive account that does not make any assumptions or claims about what the effort reflects, only that successful retrievals are more beneficial when they involve more vs. less effort). If successful retrieval is more effortful for difficult than for easy items, the REH predicts that the additional effort will benefit memory for difficult relative to easy items. Therefore, the REH predicts that final test performance may favor difficult over easy items. Furthermore, given that we were equating the numbers of correct recalls during practice (hereafter referred to as the *criterion level*), difficult items would necessarily require more trials to reach criterion, and the additional exposure might further enhance memory for the difficult items.

Only one prior study has provided evidence regarding the effects of item difficulty with equivalent criterion levels. Karpicke (2009, Exp. 1) manipulated item difficulty by collecting ease-of-learning ratings for Swahili–English pairs and then using a median split to select easy and difficult pairs. These pairs were then presented to new participants for test–restudy practice until pairs were correctly recalled one or three times. Although repeated retrieval enhanced final test performance for both difficult and easy items, recall was lower for difficult than for easy pairs (28 % vs. 41 % after one correct recall during practice, 73 % vs. 82 % after three correct recalls).

These outcomes provide preliminary evidence that item difficulty effects may persist despite equating recall success during practice, but other methodological factors limit their interpretation. First, participants practiced difficult and easy items within the same practice block during initial learning. If easy items reached criterion and dropped from practice more quickly, the remaining difficult items would be practiced with a contracting lag. Thus, differences in retention might have reflected lag effects rather than item difficulty. Second, Karpicke (2009) implemented two relatively low criterion levels (one or three correct recalls). Previous research has shown that retention improves with higher criterion levels, but that the returns diminish with increasing criterion levels (Pyc & Rawson, 2009; Vaughn & Rawson, 2011). According to the REH, these incremental gains reflect differential retrieval effort. If only the initial correct retrievals are effortful for easy items, they may quickly reach asymptote. However, if correct retrievals for difficult items continue to be effortful at higher criterion levels, we might expect additional incremental gains at higher criterion levels. Thus, we examined a wider range of criterion levels.

Experiment 1

Method

A group of 42 undergraduates participated for course credit. Item difficulty (easy or difficult) and criterion level (1, 3, 5, 7, 9, or 11 correct recalls during practice) were within-participants manipulations.

Participants learned two sets of Lithuanian–English word pairs, including 30 easy pairs and 30 difficult pairs. Items were selected on the basis of normative data (Grimaldi, Pyc, & Rawson, 2010); the normative cued recall after initial study was .40 versus .09 for our easy versus difficult items. Within each set, word pairs were randomly assigned to each of the six criterion levels (randomized anew for each participant).

Easy and difficult items were learned in two separate blocks,[1] with order (easy vs. difficult set learned first) counterbalanced across participants.[2] Each block began with a study phase, with word pairs being presented individually via computer for 10 s each. On each trial in the subsequent practice phase, a Lithuanian cue was presented, and participants had up to 8 s to type in the English target (participants could advance if they finished responding sooner). For incorrect responses, the pair was restudied for 4 s and was then placed at the end of the list for another practice trial later. For correct responses, if the item had not yet reached its assigned criterion, it was placed at the end of the list for another practice trial. Once an item had reached criterion, it was dropped from practice. After all items from one set had reached their assigned criterion, the study and practice phases for the second set were administered. Session 1 ended when all items had reached criterion or after 90 min. The data for five participants who did not learn all of the items to criterion within the time limit and for two others who did not return for the second session were excluded from the analysis.

Two days later, participants completed a final cued-recall test. The Lithuanian cues were presented one at a time, and

[1] We used separate blocks to avoid differential contracting lags for the difficult versus the easy items. Analyses of functional lag confirmed that the average lags between trials were similar for easy versus difficult items (in Exp. 1, 21.9 vs. 22.4 trials; in Exp. 2, 22.1 vs. 22.6 trials).

[2] In preliminary analyses including Item Order as a factor for both experiments, only two of 14 analyses revealed significant main effects of item order, and only one of 40 interactions was significant. Accordingly, we collapsed across counterbalancing conditions for the analyses reported in Experiments 1 and 2.

participants had unlimited time to type the English targets. After final cued recall, all 60 items were relearned. Lithuanian cues were presented one at a time in random order, and participants were prompted to retrieve the English target. After an incorrect response, the pair was restudied for 4 s and then placed at the end of the list. After a correct response, the pair was dropped from the relearning phase.

Results

Before discussing the primary measures of interest, we will report outcomes for two auxiliary measures. As a manipulation check, Table 1 reports the mean numbers of trials per item to reach criterion. A repeated measures analysis of variance (ANOVA) revealed main effects of item difficulty, $F(1, 34)=82.91$, $MSE=5.33$, $p<.001$, $\eta_p^2=.71$, and criterion level, $F(5, 170)=1{,}425.60$, $MSE=0.78$, $p<.001$, $\eta_p^2=.98$, as well as a significant interaction, $F(5, 170)=2.78$, $MSE=0.80$, $p=.019$, $\eta_p^2=.08$. Most relevant to our purposes, the difficult items required more trials to reach criterion than did the easy items.[3]

The second auxiliary measure was the first-keypress latency (i.e., the total time from cue onset to typing of the first letter) for correct responses during practice, an indicator of retrieval effort. The mean first-keypress latencies (Table 2) were longer for difficult than for easy items, $F(1, 202)=128.48$, $MSE=1.02$, $p<.001$, $\eta_p^2=.39$, confirming that successful retrieval during practice was more effortful for difficult than for easy items. Of lesser interest, the effect of criterion level was also significant, $F(5, 1010)=39.82$, $MSE=1.52$, $p<.001$, $\eta_p^2=.17$.

Concerning the primary outcomes, Fig. 1 reports final cued recall for the difficult and easy items. To revisit, a reasonable expectation was that equating the numbers of correct recalls during practice might produce equivalent final test performance for difficult and easy items; the additional trials and greater retrieval effort during practice for difficult than for easy items might even yield a performance advantage for difficult items. However, final cued recall was significantly greater for easy than for difficult items, $F(1, 34)=34.96$, $MSE=.04$, $p<.001$, $\eta_p^2=.51$ [of lesser interest, the effect of criterion level was also significant, $F(5, 170)=40.40$, $MSE=.05$, $p<.001$, $\eta_p^2=.54$]; the interaction was not significant, $F<1$. Thus, performance was consistently lower for difficult than for easy items.[4]

As a secondary measure of retention, Fig. 2 reports the mean numbers of relearning trials per item. Although all items were included in the relearning phase, we restricted analyses to those items not correctly recalled on the final cued-recall test. Providing converging evidence of a memory disadvantage for difficult items, more relearning trials were required for difficult than for easy items, $F(1, 34)=32.92$, $MSE=0.67$, $p<.001$, $\eta_p^2=.49$. The effect of criterion level, $F(5, 170)=19.48$, $MSE=0.26$, $p<.001$, $\eta_p^2=.45$, and the interaction, $F(5, 170)=5.32$, $MSE=0.22$, $p<.001$, $\eta_p^2=.06$, were also significant.

Experiment 2

Despite equating the numbers of correct recalls during practice, retention was poorer for difficult than for easy items. Why were difficult items consistently at a disadvantage?

Cued recall depends heavily on associative memory, but it also reflects target memory (and possibly cue memory). Recently, Vaughn and Rawson (2011) showed that criterion learning enhances memory for each of these components. Learners practiced Lithuanian–English word pairs via cued recall with restudy until the words were correctly recalled one to five times. Of interest here, the final test performance increased across criterion levels on cue, target, and associative recognition measures. Accordingly, the item difficulty difference in retention observed here could reflect deficits in one or more of these memory components. Experiment 2 replicated and extended Experiment 1 in order to allow us to diagnose whether the persistent disadvantage for difficult items reflected attenuated effects of criterion learning on cue memory, target memory, or associative memory.

Method

A group of 78 undergraduates participated for course credit. Item difficulty (easy or difficult) and criterion level (1, 3, 5, 7, 9, or 11 correct recalls) were within-participants manipulations. We randomly assigned participants to one of three final-test groups (cued recall, associative recognition, or cue/target recognition).

In the cue/target recognition group, participants completed both cue and target recognition tests. For cue recognition, the original 60 Lithuanian words were intermixed with 60 new Lithuanian words. For target recognition, the original 60 English words were intermixed with 60 unstudied English words. Test order (cue vs. target recognition first) was counterbalanced across participants; our analyses are

[3] Concerning other relevant outcomes, performance on the first practice trial was 45 % versus 16 % for the easy versus the difficult items. The percentages of errors during practice involving commissions for easy versus difficult items were 30 % versus 41 % in Experiment 1 and 28 % versus 35 % in Experiment 2.

[4] The percentages of errors involving commissions for easy versus difficult items were 59 % versus 61 % in Experiment 1 and 54 % versus 54 % in Experiment 2. Of these commissions, the percentages involving intralist intrusions for easy versus difficult items were 80 % versus 86 % in Experiment 1 and 79 % versus 84 % in Experiment 2.

Table 1 Mean numbers of trials to reach criterion

		Criterion Level					
Experiment	Difficulty	1	3	5	7	9	11
1	Easy	2.1 (0.1)	4.1 (0.1)	6.2 (0.1)	8.1 (0.1)	10.1 (0.2)	12.3 (0.1)
	Difficult	3.5 (0.3)	6.0 (0.3)	8.2 (0.4)	10.5 (0.4)	12.3 (0.3)	14.7 (0.3)
2	Easy	2.2 (0.1)	4.2 (0.1)	6.3 (0.1)	8.2 (0.1)	10.3 (0.1)	12.4 (0.1)
	Difficult	3.8 (0.2)	6.5 (0.3)	8.5 (0.3)	10.4 (0.2)	12.8 (0.3)	14.9 (0.3)

Standard errors of the means are reported in parentheses.

collapsed across this factor because it produced no effects on performance.

In Session 1, the materials and procedure were the same as in Experiment 1. The data for ten participants who did not learn all items to criterion within the time limit and for ten who did not return for the second session were excluded from the analysis. For Session 2, we extended the retention interval to seven days. The associative recognition group completed an associative recognition test consisting of the original 60 Lithuanian and English words, presented as 30 correctly paired Lithuanian–English words and 30 incorrectly paired Lithuanian–English words (i.e., a Lithuanian cue paired with the target from a different pair). Within each set of 30 pairs, half were easy and half were difficult (i.e., easy Lithuanian cues were always paired with easy English targets, and difficult Lithuanian cues were always paired with difficult English targets). Participants were informed that all pairs contained previously studied words and that they were to indicate whether each Lithuanian word was paired with its correct English translation.

The procedure for the final cued-recall group was the same as in Experiment 1, except that the cued-recall test was followed immediately by the associative recognition test. In all three final-test groups, items were presented one at a time in random order, and participants had unlimited time to respond. Unlike in Experiment 1, we did not administer a relearning phase.

Results

Concerning the auxiliary measures, difficult items required more trials to reach criterion than did easy items (Table 1), $F(1, 57)=85.46$, $MSE=9.76$, $p<.001$, $\eta_p^2=.60$. The effect of criterion level, $F(5, 285)=2{,}161.04$, $MSE=0.84$, $p<.001$, $\eta_p^2=.97$, and the interaction, $F(5, 285)=3.56$, $MSE=0.79$, $p=.004$, $\eta_p^2=.06$, were both significant. Additionally, first-keypress latencies for correct responses during practice were longer for difficult than for easy items (Table 2), $F(1, 326)=160.22$, $MSE=1.41$, $p<.001$, $\eta_p^2=.33$. Once again, the effect of criterion level, $F(5, 1630)=80.19$, $MSE=1.47$, $p<.001$, $\eta_p^2=.20$, and the interaction, $F(5, 1630)=2.23$, $MSE=1.21$, $p=.049$, $\eta_p^2=.01$, were significant.

In Fig. 3, we report mean cued recall. Replicating Experiment 1, performance was better for the easy than for the difficult items, $F(1, 15)=20.78$, $MSE=.03$, $p<.001$, $\eta_p^2=.58$. The effect of criterion level was also significant, $F(5, 75)=13.38$, $MSE=0.04$, $p<.001$, $\eta_p^2=.47$; the interaction was not, $F<1.4$.

For each recognition measure, performance was calculated as hits minus false alarms. To revisit, cue recognition and target recognition each involved only one set of lures; thus, corrected recognition was based on the same false alarm rate for both easy and difficult items. Associative recognition included both easy and difficult lures, and thus hits and false alarms were computed separately for easy and difficult pairs.

Table 2 Mean first-keypress latencies (in seconds) for correct responses during practice

		Criterion Level					
Experiment	Difficulty	1	3	5	7	9	11
1	Easy	2.56 (0.10)	1.98 (0.05)	1.76 (0.03)	1.66 (0.03)	1.59 (0.02)	1.51 (0.02)
	Difficult	2.82 (0.10)	2.55 (0.06)	2.21 (0.04)	2.01 (0.03)	1.92 (0.03)	1.89 (0.03)
2	Easy	2.60 (0.08)	2.09 (0.04)	1.87 (0.03)	1.63 (0.02)	1.58 (0.02)	1.53 (0.01)
	Difficult	3.02 (0.08)	2.58 (0.04)	2.24 (0.03)	2.10 (0.02)	1.97 (0.02)	1.88 (0.02)

Standard errors of the means are reported in parentheses.

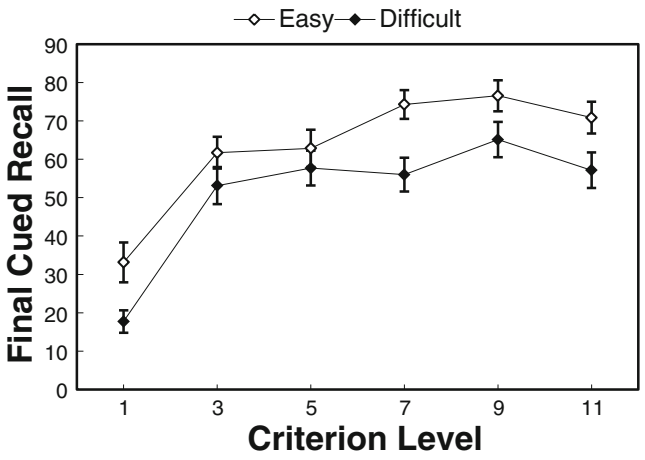

Fig. 1 Mean final cued-recall performance as a function of criterion level and item difficulty in Experiment 1. Error bars report standard errors of the means

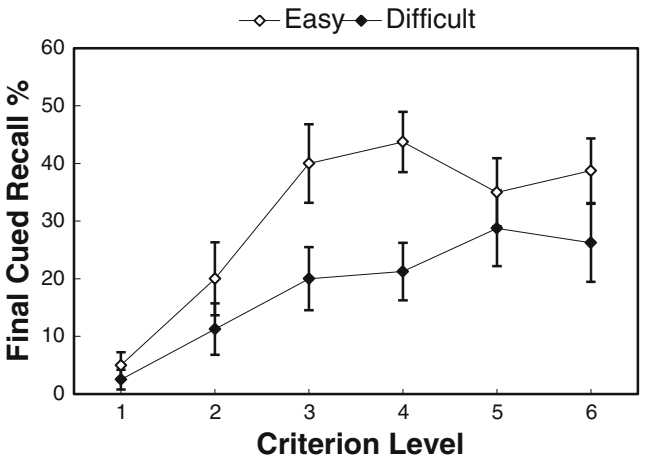

Fig. 3 Mean final cued-recall performance as a function of criterion level and item difficulty in Experiment 2. Error bars report standard errors of the means

Cue recognition did not differ significantly as a function of item difficulty (Fig. 4), $F<1.7$. The effect of criterion level was significant, $F(5, 95)=8.33$, $MSE=0.03$, $p<.001$, $\eta_p^2=.31$, but the interaction was not, $F<1.7$. Thus, the retention disadvantage for difficult relative to easy items was apparently not due to differences in cue memory.

Surprisingly, target recognition was greater for difficult than for easy targets (Fig. 5), $F(1, 19)=15.23$, $MSE=.05$, $p=.001$, $\eta_p^2=.45$. The effect of criterion level was significant, $F(5, 95)=13.51$, $MSE=0.02$, $p<.001$, $\eta_p^2=.42$, as was the interaction, $F(5, 95)=2.43$, $MSE=0.01$, $p=.04$, $\eta_p^2=.11$. Thus, the retention disadvantage for difficult items was apparently not due to differences in target memory. This measure is also the first in which performance has favored difficult over easy items. Did this advantage reflect differences in item characteristics for the target words in difficult versus easy pairs? Easy and difficult targets did not differ significantly on concreteness, imageability, or word length ($ts<0.96$). In contrast, the log word frequency (from SUBTLEXus: Brysbaert & New, 2009) was lower for difficult than for easy targets ($M=3.09$ vs. 3.46), $t(58)=2.94$, $p=.005$, $d=0.13$. Prior research has shown that recognition is better for low- than for high-frequency words (see, e.g., Reder et al., 2000). To examine whether word frequency influenced target recognition, we conducted a series of item analyses. Recognition was better for difficult than for easy targets, regardless of whether log frequency was entered as a covariate in the analyses [without frequency covariate, $F(1, 56)=26.61$, $MSE=.04$, $p<.001$, $\eta_p^2=.32$; with frequency covariate, $F(1, 55)=20.13$, $MSE=.04$, $p<.001$, $\eta_p^2=.27$], and the main effect of frequency was not significant ($F<1$). Thus, the recognition advantage for difficult targets does not appear to reflect word frequency. For completeness, we also conducted item analyses for cued recall. Cued

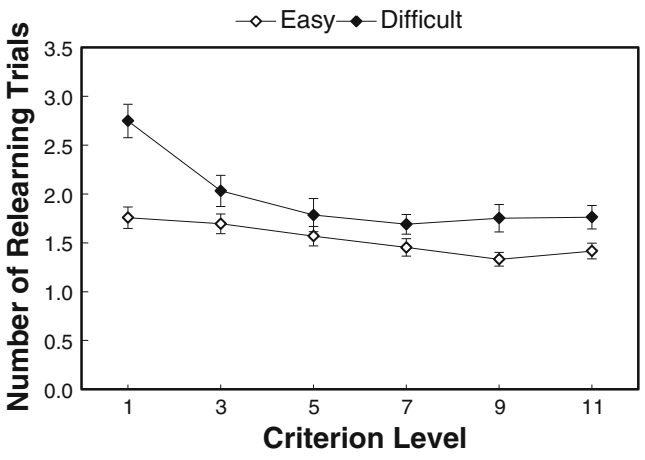

Fig. 2 For items that were not correctly recalled on the final cued-recall test, mean numbers of relearning trials per item as a function of criterion level and item difficulty in Experiment 1. Error bars report standard errors of the means

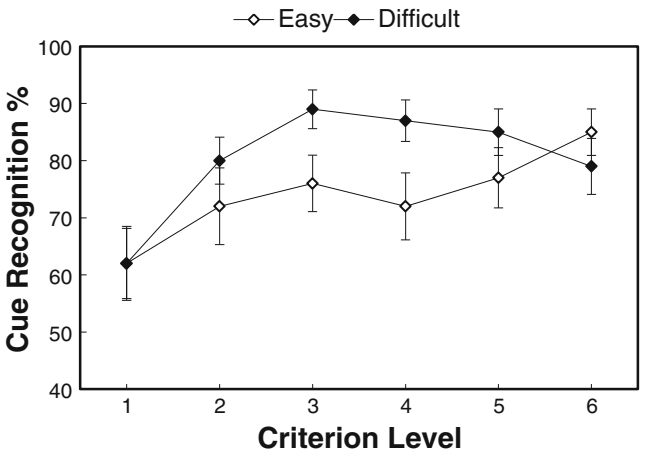

Fig. 4 Mean final cue recognition performance as a function of criterion level and item difficulty in Experiment 2. Error bars report standard errors of the means

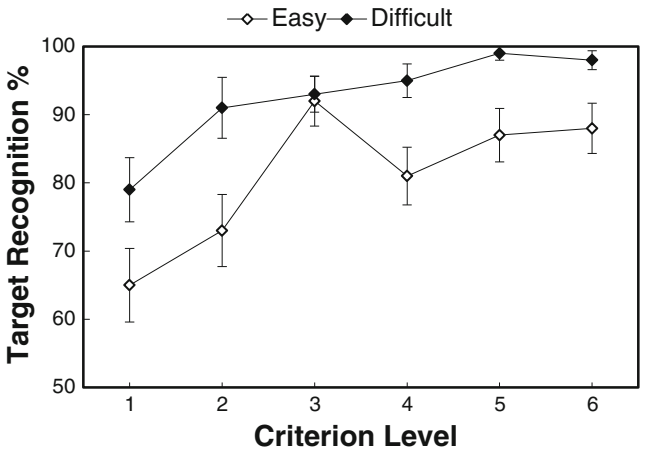

Fig. 5 Mean final target recognition performance as a function of criterion level and item difficulty in Experiment 2. Error bars report standard errors of the means

recall was better for easy than for difficult word pairs, regardless of whether log frequency was entered as a covariate [without covariate, $F(1, 47)=12.37$, $MSE=.08$, $p=.001$, $\eta_p^2=.21$; with covariate, $F(1, 46)=9.65$, $MSE=.08$, $p=.003$, $\eta_p^2=.17$]; the effect of frequency was not significant ($F<1$).

Concerning associative recognition, performance did not differ for participants who did or did not complete cued recall prior to the associative recognition test ($F<1$); therefore, we collapsed across groups. Corrected associative recognition was greater for easy than for difficult items (65 % vs. 54 %), $t(37)=3.05$, $p=.004$, $d=1.00$. This advantage was due primarily to fewer false alarms for easy than for difficult pairs (22 % vs. 30 %), $t(37)=3.43$, $p=.001$, $d=1.13$; the numerical trend for greater hits to easy than to difficult pairs (88 % vs. 85 %) was not significant, $t(37)=1.14$, $p=.260$, $d=0.37$. Thus, the cued-recall disadvantage for difficult relative to easy items most likely is due to poorer associative memory.

General discussion

Despite equating the numbers of successful recalls during practice, subsequent cued recall was consistently poorer for difficult than for easy items. Experiment 2 indicated that the retention deficit for difficult items was due primarily to poorer associative memory. Although successful retrieval is a potent memory enhancer, it cannot overcome differences in item difficulty when difficult and easy items are recalled the same number of times during practice.

Concerning the theoretical implications, the persistent item difficulty effect in cued recall is inconsistent with patterns that would reasonably be predicted according to the REH, given that successful retrieval was more effortful for difficult items. Notably, the REH is a descriptive account that is silent on what the additional effort of successful retrieval might reflect, and thus has limited explanatory power. Other theories have suggested that in some cases, effort may reflect an elaborative search through memory, which enhances memory (i.e., the elaborative retrieval hypothesis; Carpenter, 2011). However, the additional effort for difficult items here did not likely reflect elaborative retrieval, given the lack of a performance advantage. Thus, the REH is limited without further clarification of the underlying processes that additional effort might reflect.

With that said, one might argue that the predictions of the REH do not apply to situations in which different sets of items are used (e.g., easy vs. difficult items), but only to different conditions that affect retrieval effort. Of relevance here, prior research has equated targets but manipulated cue effectiveness to make retrieval more versus less difficult (e.g., Carpenter & DeLosh, 2006; Finley, Benjamin, Hays, Bjork, & Kornell, 2011). Note that this logic rests on equivalent targets. Given that the easy and difficult word pairs used here consisted of nominally different target words, we collected normative data to establish that our target words were functionally equivalent. In brief, we split our materials into three sets of 20 (each with ten easy and ten difficult pairs). Each participant studied one set (10 s per item), and after a 5-min distractor task, completed free recall of either the Lithuanian cues or English targets. Free recall was better for easy than for difficult cues (19.1 % vs. 5.9 %), $t(57)=7.46$, $p<.001$, although note that this measure of cue memory does not directly assess cue effectiveness (i.e., how effectively the cue word elicits the associated target word). More importantly, free recall did not differ for the easy versus difficult targets (49.1 % vs. 48.4 %), $t(56)=0.28$, $p=.782$, suggesting functional equivalence of the two sets of target words. Thus, these normative data suggest that the cued-recall disadvantage for difficult versus easy word pairs is not due to a priori differences in the memorability of target words.

Rather, the associative recognition outcomes in Experiment 2 point to a deficit in associative memory for the difficult versus the easy word pairs. Why did we find a deficit in associative memory for the difficult pairs? One factor that contributes to associative memory is the use of mediators, such as the use of keywords to link the cue and target words. For a mediator to be effective, the cue needs to elicit the mediator (i.e., *mediator retrieval*) and the recalled mediator needs to elicit the correct target (i.e., *mediator decoding*). Given that effective mediators enhance subsequent associative memory (e.g., Dunlosky, Hertzog, & Powell-Moman, 2005), one potential explanation for the pattern observed here is that less effective mediators were used for difficult than for easy items. Although the present work was not designed to directly explore this possible explanation, this account could be directly tested by

including previously established measures of mediator retrieval and mediator decoding (e.g., Dunlosky et al., 2005; Pyc & Rawson, 2010). For example, Pyc and Rawson (2010) had participants report their keyword mediators during learning of word pairs. On a subsequent final test, mediator retrieval was assessed by presenting participants with cue words and prompting them to recall the mediator that they had generated during practice. To assess mediator decoding, another group was presented with both the cue word and the corresponding mediator that the participant had generated during practice, and participants were prompted to recall the target. If these measures were added to the present design, a testable prediction would be that performance on these measures of mediator retrieval and mediator decoding would be greater for easy than for difficult word pairs. This account points to a fruitful direction for further exploration of why item difficulty differences linger, even after equating recall success during practice.

Finally, the present results have practical implications for optimizing study schedules: Students should be made aware that, although repeated retrieval enhances memory performance in general, equating criterion levels during learning will not equate final performance for difficult versus easy items, and that additional retrieval practice and/or the use of additional encoding strategies may be needed for students to achieve their desired level of mastery for difficult items.

Author Note The research reported here was supported by a James S. McDonnell Foundation 21st Century Science Initiative in Bridging Brain, Mind and Behavior Collaborative Award.

References

Bahrick, H. P., & Hall, L. K. (2005). The importance of retrieval failures to long-term retention: A metacognitive explanation of the spacing effect. *Journal of Memory and Language, 52*, 566–577. doi:10.1016/j.jml.2005.01.012

Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods, 41*, 977–990. doi:10.3758/BRM.41.4.977

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

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

Cull, W. L., & Zechmeister, E. B. (1994). The learning ability paradox in adult metamemory research: Where are the metamemory differences between good and poor learners? *Memory & Cognition, 22*, 249–257.

Dunlosky, J., Hertzog, C., & Powell-Moman, A. (2005). The contribution of five mediator-based deficiencies to age-related differences in associative learning. *Developmental Psychology, 41*, 389–400.

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

Grimaldi, P. J., Pyc, M. A., & Rawson, K. A. (2010). Normative multitrial recall performance, metacognitive judgments, and retrieval latencies for Lithuanian–English paired associates. *Behavior Research Methods, 42*, 634–642. doi:10.3758/BRM.42.3.634

Karpicke, J. D. (2009). Metacognitive control and strategy selection: Deciding to practice retrieval during learning. *Journal of Experimental Psychology. General, 138*, 469–486. doi:10.1037/a0017341

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

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

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

Reder, L. M., Nhouyvanisvong, A., Schunn, C. D., Ayers, M. S., Angstadt, P., & Hiraki, K. (2000). A mechanistic account of the mirror effect for word frequency: A computational model of remember–know judgments in a continuous recognition paradigm. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26*, 294–320. doi:10.1037/0278-7393.26.2.294

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

Wissman, K. T., Rawson, K. A., & Pyc, M. A. (2012). How and when do students use flashcards? *Memory, 20*, 568–579. doi:10.1080/09658211.2012.687052