Mem Cogn (2012) 40:976–988
DOI 10.3758/s13421-012-0200-x

Are judgments of learning made after correct responses during retrieval practice sensitive to lag and criterion level effects?

Mary A. Pyc · Katherine A. Rawson

Published online: 8 March 2012

Abstract Although successful retrieval practice is beneficial for memory, various factors (e.g., lag and criterion level) moderate this benefit. Accordingly, the efficacy of retrieval practice depends on how students use retrieval practice during learning, which in turn depends on accurate metacognitive monitoring. The present experiments evaluated the extent to which judgments of learning (JOLs) made after correct responses are sensitive to factors (i.e., lag and criterion level) that moderate retrieval practice effects, as well as which cues influence JOLs under these conditions. Participants completed retrieval practice for word pairs with either short or long lags between practice trials until items were correctly recalled 1, 3, 6, or 9 times. After the criterion trial for an item, participants judged the likelihood of recalling that item on the final test 1 week later. JOLs showed correct directional sensitivity to criterion level, with both final test performance and JOLs increasing as criterion level increased. However, JOLs showed incorrect directional sensitivity to lag, with greater performance but lower JOLs for longer versus shorter lags. Additionally, results indicated that retrieval fluency and metacognitive beliefs about criterion level—but not lag—influenced JOLs.

Keywords Metamemory · Memory · Recall

M. A. Pyc (✉)
Department of Psychology, Washington University in St Louis,
Box 1125, St Louis, MO 63130, USA
e-mail: mpyc@wustl.edu

K. A. Rawson
Department of Psychology, Kent State University,
P.O. Box 5190, Kent, OH 44242, USA

A wealth of research has shown that practice involving retrieval of target information from memory (i.e., retrieval practice) is beneficial for subsequent retention (for reviews, see Rawson & Dunlosky, 2011; Roediger & Butler, 2011). Of course, the effectiveness of retrieval practice depends on a number of factors. For example, although failed retrieval attempts may show modest memorial benefits (e.g., Kornell, Hays, & Bjork, 2009), retrieval practice is particularly efficacious when retrieval attempts during encoding are successful (e.g., Karpicke & Roediger, 2007; Pyc & Rawson, 2007, 2011). Furthermore, the memorial benefits of successful retrievals depend critically on the quantity and timing of those successful retrievals (Pyc & Rawson, 2009).

Although retrieval practice has been shown to yield large improvements in memory under appropriate experimentally devised conditions, in many learning situations (e.g., a student studying for an exam), the scheduling of retrieval practice is largely in the hands of the learner. Thus, the efficacy of retrieval practice for enhancing learning can only be as good as individuals' self-regulated use of retrieval practice. Therefore, it is important to understand the extent to which individuals' judgments of learning are sensitive to factors that influence the efficacy of retrieval practice. Accordingly, the present research examined the extent to which individuals' judgments are sensitive to the quantity and timing of successful retrievals during practice.

Below, we provide a brief review of the particular retrieval practice effects that are relevant for the present experiments. We then describe components of self-regulated learning, with particular emphasis on metacognitive monitoring, the component of greatest interest here. Finally, we report two experiments evaluating the sensitivity of judgments of learning to factors that influence the efficacy of successful retrieval practice.

Efficacy of retrieval practice

Many studies have established that retrieval practice is beneficial for memory. Retrieving information from memory during practice promotes memory to a greater extent than do other strategies, such as restudying (e.g., Cull, 2000; Karpicke & Roediger, 2007, 2008). Important for present purposes, previous research has shown that the quantity and timing of practice influences the memorial benefits of retrieval practice.

Concerning the quantity of practice, research has shown greater memorial benefits when individuals engage in more versus less retrieval during practice (e.g., Allen, Mahler, & Estes, 1969; Wheeler & Roediger, 1992). Concerning the timing of practice, a wealth of previous research has demonstrated greater memorial benefits when items are practiced with a longer versus shorter lag between practice trials with items (e.g., Cepeda, Pashler, Vul, Wixted, & Rohrer, 2006; Cull, 2000; Landauer & Bjork, 1978; Pashler, Zarrow, & Triplett, 2003; Pyc & Rawson, 2009). However, almost all of this previous research has manipulated the quantity and timing of *trials* during practice. In contrast, the present research involved manipulating the quantity and timing of *correct retrievals* during practice. When students self-regulate their own learning using retrieval practice, they presumably do not (and should not) simply engage in a fixed number of practice trials for each item. Rather, students should self-test until they can correctly recall items multiple times during encoding (e.g., Pyc & Rawson, 2009).

What influence does the quantity and timing of correct retrievals have on final test performance? Recent research has shown greater memorial benefits for items correctly retrieved more versus fewer times during practice and for items that are correctly retrieved after longer versus shorter lags during retrieval practice (Pyc & Rawson, 2009). Pyc and Rawson (2009) presented participants with foreign language paired associates for an initial study trial and then test–restudy practice trials until items reached a preassigned criterion level of performance (1, 3, 5, 6, 7, 8, or 10 correct retrievals) during practice. Items were practiced with either a short or a long lag between practice trials. After a delay, participants completed a final cued recall test for all items. Across two experiments, performance increased as the number of correct retrievals during practice increased (see also Nelson, Leonesio, Shimamura, Landwehr, & Narens, 1982; Vaughn & Rawson, 2011). Additionally, performance was higher for items with a longer lag versus shorter lag between correct retrievals during practice. Thus, the benefits of successful retrievals depend critically on the quantity and timing of those successful retrievals.

Theories of self-regulated learning and metacognitive monitoring

Although researchers have identified various retrieval practice schedules that are particularly beneficial for memory (i.e., schedules with multiple correct retrievals that take place after long lags), the impact of successful retrieval practice for promoting learning hinges critically on individuals' using the most effective retrieval practice schedules when self-regulating their study. Self-regulated learning includes two central components, *monitoring* and *control* (e.g., Greene & Azevedo, 2007; Nelson & Narens, 1990; Winne & Hadwin, 1998). Monitoring involves evaluating how well information has been learned and/or the likelihood that information will be remembered in the future. Control involves decisions about what to study, when to study, and how to study. The primary assumption of models of self-regulated learning is that monitoring informs control decisions, which in turn influence learning (e.g., Ariel, Dunlosky, & Bailey, 2009; Dunlosky & Metcalfe, 2009; Nelson & Narens, 1990; Winne & Hadwin, 1998). Consistent with this basic assumption, research has shown that more accurate versus less accurate monitoring during study leads to higher levels of test performance (e.g., Dunlosky & Rawson, in press; Rawson, O'Neil, & Dunlosky, 2011; Thiede, 1999; Thiede, Anderson, & Therriault, 2003).

Because monitoring accuracy is critically important for effective control and later test performance, we focus on this aspect of self-regulated learning in the present experiments. To examine the extent to which individuals accurately monitor their learning during retrieval practice, we evaluated the extent to which judgments of learning (JOLs) made after correct retrievals are sensitive to factors (i.e., lag and criterion level) that moderate the effects of successful retrieval.

What factors influence JOLs? Koriat's (1997) cue-utilization framework states that JOLs are inferential, in that individuals do not have direct access to their own memory states and, thus, must use heuristics to assess the likelihood of being able to later recall information. That is, JOLs are not based on an evaluation of the memory strength of an item but, instead, are based on one or more cues that individuals use to infer the state of their memory.

What types of cues are used to make JOLs? According to the cue-utilization framework, three classes of cues can influence JOLs: intrinsic, extrinsic, and mnemonic. Intrinsic cues are based on characteristics inherent to items, which may make them easier or more difficult to learn (e.g., abstract vs. concrete). Extrinsic cues are based on learning conditions (e.g., number of trials) or the encoding task an individual engages in (e.g., interactive imagery). Mnemonic cues are based on aspects of an individual's own subjective experiences during task performance (e.g., retrieval fluency), which may provide the individual with information that is predictive of how well an item has been learned, as well as the likelihood that the item will be recalled at a later time. To foreshadow, extrinsic and mnemonic cues are of greatest interest here.

Sensitivity of JOLs to effects of correct retrievals

With the goal of the present research in mind (i.e., to evaluate the sensitivity of JOLs to the quantity and timing of successful retrievals during practice), to what extent can previous research provide information about the kinds of cues that learners use to make JOLs after correct retrievals?

A wealth of previous research has evaluated the sensitivity of JOLs to the quantity and timing of practice, but these previous studies are different in important ways from the present research. For example, previous research has shown greater JOL accuracy as the quantity of practice increases (e.g., Mazzoni, Cornoldi, & Marchitelli, 1990; Meeter & Nelson, 2003; Zechmeister & Shaughnessy, 1980). However, much of this previous research has involved study trials only, rather than retrieval practice. Furthermore, prior research involving retrieval practice has manipulated the number of practice trials, rather than manipulating the number of correct retrievals.

Likewise, previous research has examined JOL accuracy as a function of timing of practice. JOL magnitudes are often greater with less versus more time between practice trials, whereas performance is usually lower with less versus more time between practice trials (e.g., Kornell, 2009; Zechmeister & Shaughnessy, 1980). However, the available research either has again involved only study trials or has manipulated the timing of practice trials rather than the timing of correct retrievals. Furthermore, much of the work showing JOL magnitude differences as a function of timing has compared massed versus spaced practice (i.e., no spacing vs. some spacing between practice trials with items), rather than short versus long lags.

Why are these differences important? First, given that previous research has shown differences in JOL accuracy for study versus retrieval practice (e.g., Karpicke, 2009; Kornell & Son, 2009; Mazzoni & Nelson, 1995, Experiment 2; Shaughnessy & Zechmeister, 1992), the sensitivity of JOLs to effects of the quantity and timing of practice in previous studies involving study trials only may differ from the sensitivity of JOLs to these factors under conditions involving retrieval practice (e.g., because the mnemonic cue of retrieval fluency is available under conditions of retrieval practice, but not under conditions of study only). Second, implementing a fixed number of practice *trials* for each item yields differences in learning status for various items. That is, some items may be correctly recalled during practice, whereas others may not be correctly recalled. Because retrieval status (i.e., correct vs. incorrect) is a powerful cue for making judgments (Nelson & Dunlosky 1991), differences in retrieval status for individual items exerts a strong influence on JOLs made during practice with a fixed number of trials. In contrast, when all items are learned to a given criterion, individuals cannot use retrieval status as a cue for making judgments. Third, a similar logic applies to studies manipulating the lag between trials, rather than the lag between correct retrievals, in that retrieval status will differ as a function of lag in the former case, but not in the latter. In sum, the sensitivity of JOLs to the quantity and timing of correct retrievals may differ from patterns observed in previous research to the extent that the available cues differ for conditions of criterion versus noncriterion learning.

Importantly, here we are interested in students' judgments of learning when all items are successfully retrieved, for reasons described above. However, to our knowledge, only one prior study has examined JOLs during criterion learning (i.e., when all items are practiced until correctly recalled). Karpicke (2009) reported that JOLs were greater for items that were correctly recalled three versus one time during practice.[1] No prior research has evaluated the relationship between lag and JOLs when items are learned to a criterion level of performance, nor has prior research examined JOLs when both lag and criterion level are manipulated.

However, on the basis of the kinds of cues that Koriat's (1997) cue-utilization framework assumes people use when making JOLs, we outline a number of possible outcomes. On the basis of the definition provided by the cue-utilization framework, criterion level is an extrinsic cue. If individuals have accurate beliefs regarding criterion level, JOLs will increase as criterion level increases. Of course, even if participants have accurate beliefs, it is possible that they may not use these beliefs when making JOLs (e.g., Koriat, Bjork, Sheffer, & Bar, 2004), so one might not see a relationship between criterion level and JOL. It could also be

[1] One other earlier study did not collect JOLs but did report parallel results using a related kind of judgment. Leonesio and Nelson (1990) had participants learn items to a criterion of one or four correct recalls. After the learning phase, all items appeared on the screen, and participants were asked to rank the items on the basis of how well they believed that they knew them. Judgments of knowing (JOKs) were greater for items that were correctly recalled four versus one time during practice. However, given that items were presented in an array format during the judgment phase of the study, when ranking items, participants may have made item-to-item comparisons that differ from the bases of sequential JOLs (cf. Thiede & Dunlosky, 1999). Furthermore, JOKs do not include the predictive component that requires consideration of forgetting, as do JOLs (for evidence that these two kinds of judgments differentially reflect estimates of forgetting, see Rawson, Dunlosky, & McDonald, 2002).

the case that individuals do not have any beliefs about criterion level, in which case JOLs will not differ for various criterion levels. (We do not consider the highly implausible possibility that individuals would believe that an increase in criterion level would lead to a decrease in memory.)

Although criterion level is an extrinsic cue, it also influences the mnemonic cue of retrieval fluency. For example, metacognitive research has shown that in various tasks, JOLs increase as response latencies decrease (e.g., Benjamin, Bjork, & Schwartz, 1998). Importantly, previous research on retrieval practice has shown that retrieval latencies decrease as the number of correct retrievals during practice increases (e.g., Pyc & Rawson, 2009). Therefore, if JOLs during retrieval practice are based on the mnemonic cue of retrieval fluency, JOLs are predicted to increase as criterion level increases.

Lag is also an extrinsic cue by definition. If individuals have accurate beliefs about lag, JOLs will be higher for items that are correctly retrieved after longer versus shorter lags. Again, even if participants have accurate beliefs, it does not ensure that they will use these beliefs when making JOLs (Koriat et al., 2004), in which case JOLs may not differ for longer versus shorter lags. JOLs also may not be related to lag if individuals do not have any beliefs about the effects of lag. Finally, if individuals have inaccurate beliefs about lag (and incorporate those beliefs when making JOLs), JOLs will be higher for shorter versus longer lags.

The extrinsic cue of lag also influences the mnemonic cue of retrieval fluency. Previous research has shown that retrieval latencies during retrieval practice are lower for items retrieved after shorter versus longer lags (e.g., Pyc & Rawson, 2009). If JOLs during retrieval practice are based on the mnemonic cue of retrieval fluency, JOLs will be higher for items that are correctly retrieved after shorter versus longer lags.

The present experiments were designed to evaluate two questions. First, are JOLs sensitive to the effects of criterion level and/or the lag between correct retrievals on final test performance? Second, what cues are used to make JOLs for criterion level and lag? In two experiments, participants learned foreign language paired associates via retrieval practice with restudy until items reached an assigned criterion level of performance (one, three, six, or nine correct retrievals). Items were practiced with either a short lag or a long lag between trials. After the last correct retrieval for each item, participants predicted the likelihood of retrieving that item on the final test. If JOLs are based on the extrinsic cue of criterion level and/or on the mnemonic cue of retrieval fluency, JOLs will increase as criterion level increases. For lag, several outcomes are plausible, depending on the extent to which the extrinsic cue of lag complements or competes with the mnemonic cue of retrieval fluency.

Experiment 1

Method

Participants and design Forty-one Kent State University undergraduates participated in return for course credit. Criterion level (one, three, six, or nine correct retrievals during practice) was a within-participants manipulation. Lag (short vs. long) was a between-participants manipulation, with 22 and 19 participants in each group, respectively.

Materials Items included 48 Swahili–English translation word pairs previously normed for item difficulty (Nelson & Dunlosky, 1994). Twelve word pairs were assigned to each of four lists, with an equivalent range of item difficulty in each list. Within each list, three items were randomly assigned to each criterion level (randomized anew for each participant).

Procedure All task instructions and items were presented via computer. All items first received an initial study trial, followed by blocks of test–restudy practice trials until items reached their assigned criterion level of performance. For initial study trials, the cue (Swahili word) and target (English translation) appeared on the computer screen for 10 s. For test trials, the cue appeared on the computer screen, and participants had 8 s to type the correct target answer. If an item was retrieved before 8 s had elapsed, participants could press a key to submit their response. Items that were not correctly retrieved received a 4-s restudy trial with the cue and target before participants moved on to the next to-be-learned item. Items that were correctly retrieved did not receive a restudy trial before participants moved on to the next item.

The computer tracked the number of times each item was correctly retrieved during practice. Items continued to receive test–restudy practice trials until they reached their assigned criterion level of performance (one, three, six, or nine correct retrievals). After items reached their criterion level of performance, they were dropped from further test–restudy practice. If an item had not reached its criterion level of performance on a given trial, it was placed at the end of the list of to-be-learned items. Participants were not aware of the specific criterion level for each item but were told that items would be practiced until they reached an “acceptable level of performance.”

For the short-lag group, the 12 items from one list were each presented for an initial study trial. After all items in the list had an initial study trial, items received test–restudy practice trials until they were correctly retrieved to their predetermined criterion level. When all items in one list had been practiced to criterion, items from a second list were presented for initial study and test–restudy practice trials, and so on until items from each of the four lists had been learned. Order of list presentation was counterbalanced across participants.

For the long-lag group, the four lists of 12 items were combined into one list. All items were presented for an initial study trial. After initial study, items received test–restudy practice trials until items were correctly retrieved to their predetermined criterion level.

Immediately after a given item was correctly recalled to its criterion level of performance (i.e., one, three, six, or nine correct retrievals), participants made a JOL for that item. For the JOL trial, participants were asked the following: "For the item you just saw, how likely do you think it is that you will be able to correctly recall the ENGLISH translation when you are shown the SWAHILI word on the final test 7 days from now?" Participants were asked to type in a response, using any number from 0 to 100 (in which 0 = *0% likelihood of recalling in 7 days* and 100 = *100% likelihood of correctly recalling in 7 days*). Thus, participants made 48 JOLs, one for each item immediately after the item reached its criterion level of performance during practice.

During the second session 1 week later, participants completed a computer -administered self-paced cued recall final test for all 48 word pairs.

Results and discussion

Final test performance The mean percentage of items correctly recalled on the final test as a function of criterion level and lag is presented in Fig. 1. Results of a 2 (lag) × 4 (criterion level) mixed factor analysis of variance (ANOVA) showed a significant main effect of criterion level, with final test performance significantly increasing as the number of correct retrievals during practice increased, $F(3, 117) = 41.67$, $MSE = .02$, $p < .001$. The main effect of lag was also significant, with final test performance significantly higher in the long-lag group than in the short-lag group, $F(1, 39) = 37.74$, $MSE = .07$, $p < .001$. The interaction was also significant, indicating a greater difference in performance for the lag groups as criterion level increased, $F(3, 117) = 6.47$, $MSE = .02$, $p < .001$.

Judgments of learning As was expected on the basis of findings from prior research, higher criterion levels and longer lags between correct retrievals improved final test performance. More important for present purposes, to what extent were JOLs sensitive to the effects of criterion level and lag on final test performance? Mean JOL values at each criterion level for each lag group are presented in Fig. 2. Results of a 2 (lag) × 4 (criterion level) mixed factor ANOVA showed a significant main effect of criterion level, with mean JOL values increasing as the number of correct retrievals during practice increased, $F(3, 117) = 44.16$, $MSE = 162.78$, $p < .001$. Thus, JOLs show correct directional sensitivity to the effects of criterion level on final test performance.

In contrast, the main effect of lag was not significant, $F(1, 39) = 2.26$, $MSE = 2{,}942.23$, $p = .141$. JOLs did not accurately reflect the effects of lag on final test performance. In fact, the numerical trend was in the opposite direction (*t*-tests showed a significant difference between short-lag and long-lag JOLs for criterion level 1, $t(39) = 2.61$, $p = .01$, as well as a trend for criterion level 3, $t(39) = 1.81$, $p = .08$). The interaction term was not significant, $F(3, 117) = 2.12$, $MSE = 162.78$, $p = .102$. Thus, although performance differences between lag groups increased as criterion level increased, JOL differences did not show this same pattern.

In sum, JOLs showed correct directional sensitivity to the effects of criterion level but did not show correct directional sensitivity to the effects of lag between correct retrievals. To what extent did the mnemonic cue of retrieval fluency influence JOLs? To measure retrieval fluency, we examined first keypress latency for all correct retrieval trials in session 1. First keypress latency was defined as the amount of time between onset of the Swahili cue and a participant's first

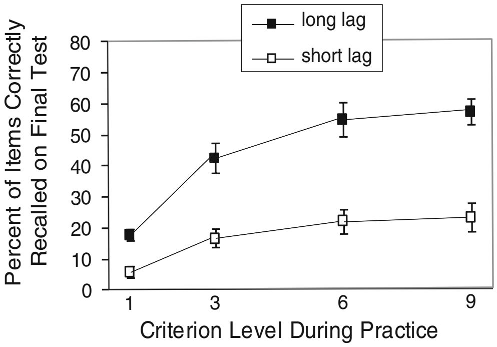

Fig. 1 Mean percentage of items correctly recalled on the final test as a function of criterion level and lag, Experiment 1. Error bars represent standard errors

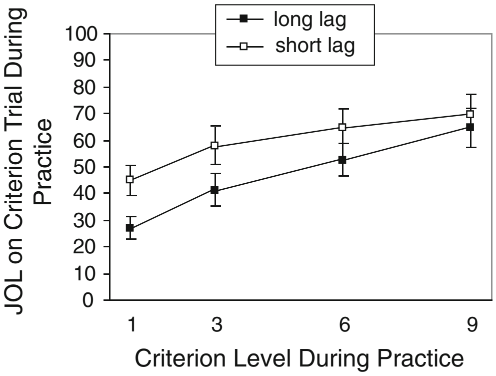

Fig. 2 Mean JOL values on criterion trial during practice as a function of criterion level and lag, Experiment 1. Error bars represent standard errors

keypress in the response box. For each participant, we calculated the mean first keypress latency for the *n*th correct retrieval during practice, with n = 1–9 correct retrievals across criterion level conditions. To provide the most stable estimates of first keypress latency, we collapsed across criterion level for this analyses (e.g., all 48 items were correctly recalled once and thus contributed to this mean, the 36 items assigned to criterion levels 3–9 were each correctly recalled a second and third time and thus contributed to these means, and so on; outcomes were highly similar when analyses were conducted only on the basis of items assigned to criterion 9). Figure 3 shows mean first keypress latency (in seconds) as a function of the *n*th correct retrieval during practice. Results of a 2 (lag) × 9 (*n*th correct retrieval) mixed factor ANOVA revealed a significant main effect of lag, with shorter latencies for the short-lag group than for the long-lag group, $F(1, 39) = 8.56$, $MSE = .50$, $p = .006$. The main effect of *n*th correct retrieval during practice was also significant, with latencies decreasing as the number of correct retrievals during practice increased, $F(8, 312) = 213.57$, $MSE = .04$, $p < .001$. The interaction was also significant, $F(8, 312) = 7.83$, $MSE = .04$, $p < .001$.

These results support the possibility that the mnemonic cue of retrieval fluency influenced JOLs during criterion learning. However, at least for criterion level, the extrinsic cue may also have influenced JOLs. Given that both mnemonic and extrinsic cues may influence JOLs, we examined the extent to which criterion level and retrieval fluency uniquely influence JOLs by conducting a series of hierarchical linear models (HLMs).[2] We also examined the extent to which two other cues may have influenced JOLs. Specifically, we included the intrinsic cue of normative item difficulty (from Nelson & Dunlosky, 1994) and the mnemonic cue of number trials involving retrieval failure prior to the first correct recall during practice for each item. The first model assessed the relationship between criterion level and JOLs. Results showed that JOLs significantly increased as criterion level increased, $t(1926) = 7.20$, $p < .001$. The second model assessed the relationship between retrieval fluency (first keypress latency) and JOLs. Results showed that JOLs significantly increased as first keypress latencies decreased, $t(1926) = 7.03$, $p < .001$. The third and fourth models assessed the relationship between normative item difficulty and JOLs and between number of retrieval failures and JOLs, respectively. Results showed no significant relationship between either of these variables and JOLs, *p*s > .05.

Given the significant relationships between criterion level and JOLs and retrieval fluency and JOLs, the fifth model examined the extent to which each of these variables influenced JOLs when the other variable was controlled for. Results showed that both criterion level and first keypress latency were significantly related to JOLs, $t(1925) = 6.52$, $p < .001$, and $t(1925) = 2.16$, $p = .03$, respectively. Taken together, these analyses suggest that both the factors of criterion level and retrieval fluency influenced JOLs during retrieval practice.

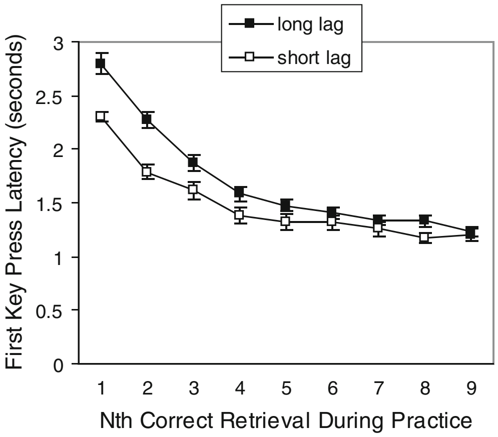

Fig. 3 Mean first keypress latency (in seconds) as a function of the *n*th correct retrieval during practice for each lag group, Experiment 1. Error bars represent standard errors

Experiment 2

Results demonstrated that JOLs show correct directional sensitivity to the effects of criterion level on final test performance: Both final test performance and JOLs increased as criterion level increased. Furthermore, HLM analyses indicated a relationship between the extrinsic cue of criterion level and JOLs above and beyond the influence of criterion level on the mnemonic cue of retrieval fluency. Presumably, the extrinsic cue reflects a metacognitive belief about the effects of criterion level on final test performance. However, Karpicke (2009) reported results suggesting that learners may not have appropriate metacognitive beliefs regarding criterion level. Of interest here, after items were learned to criterion during practice, participants were asked to make aggregate judgments, in which they judged the number of items they would remember on a final test 1 week later. Results showed that aggregate judgments did not differ for a group of participants who terminated practice after one correct recall versus participants who completed two additional practice trials, suggesting that participants may not understand the memorial benefits of increasing criterion levels. Thus, one goal of Experiment 2 was to provide further evidence that participants have correct metacognitive beliefs about the effects of criterion level on final test performance.

[2] For an explanation for conditions under which use of multilevel models are appropriate, see Schwartz and Stone (1998).

In contrast to the criterion level results, JOLs did not show correct directional sensitivity to the effects of lag between correct retrievals on final test performance. Final test performance was higher for the long-lag versus short-lag group, whereas JOLs did not statistically differ (and were even numerically lower) for the long-lag versus short-lag group. The design of Experiment 1 precluded us from examining the relationship between lag and JOLs using HLM analyses, as we did for criterion level, because lag was a between-participants manipulation. Therefore, in Experiment 2, lag was manipulated within subjects. Additionally, to further diagnose why JOLs did not show correct directional sensitivity to the effects of lag on final test performance, Experiment 2 evaluated metacognitive beliefs about the effects of lag. One possibility is that participants have correct metacognitive beliefs about the effects of lag on final test performance, but the salient mnemonic cue of retrieval fluency overrides the extrinsic cue of lag. Another possibility is that participants do not have beliefs or have incorrect beliefs about the effects of lag. To measure participants' metacognitive beliefs about the effects of criterion level and lag on final test performance, in addition to making item-specific JOLs, participants in Experiment 2 also made *aggregate judgments*. In contrast to item-specific JOLs, aggregate judgments are global predictions about performance, in which participants make overall judgments about the number of items within each level of lag and criterion they believed they will later recall.

The results of Experiment 1 are consistent with the idea that participants have correct beliefs about the effects of criterion level on final test performance, and thus we predicted that aggregate judgments would be greater for higher versus lower criterion levels. In contrast, the pattern of results for lag will be more revealing because a number of outcomes are plausible. If participants have correct beliefs about the effects of lag on final test performance, aggregate judgments will be greater for longer versus shorter lags. If participants do not have beliefs about the effects of lag on final test performance, aggregate judgments will not differ for longer versus shorter lags. Finally, if participants have incorrect beliefs about the effects of lag on final test performance, aggregate judgments will be greater for shorter versus longer lags.

Method

Participants and design Sixty-seven Kent State University undergraduates participated in return for course credit. Criterion level (one, three, or nine correct retrievals per item) and lag (short vs. long) were within-participants manipulations. Aggregate judgment (preacquisition versus no preacquisition judgment) was a between-participants manipulation, with 36 and 31 participants in each group, respectively.

Materials Items included 72 Swahili–English translation word pairs. Thirty-six word pairs were assigned to each of two lists, with an equivalent range of item difficulty within each list. List assignment to lag was counterbalanced across participants. Twelve items within each list were randomly assigned to each criterion level (randomized anew for each participant).

Procedure All participants were told that they would be learning foreign language word pairs and would receive test–restudy practice with items until they reached an acceptable level of performance. Participants in the *preacquisition aggregate judgment* group then received detailed instructions about lag and criterion level (see the Appendix for complete instructions). In brief, we described these variables in relation to studying with flashcards, a common study strategy reported by undergraduates (Kornell & Bjork, 2008). Participants then predicted how many short-lag and long-lag items they would be able to remember on the final test. Next, they predicted how many of the criterion 1, 3, and 9 items they would be able to remember. These preacquisition aggregate judgments were included as a measure of prior knowledge regarding lag and criterion level.

All participants then began the main experimental task. As in Experiment 1, items were each presented for a 10-s initial study trial. After initial study, items received test–restudy practice until they were correctly retrieved to their criterion level of performance. Once an item reached criterion, participants made a JOL for the item, and then the item was dropped from the list. For short-lag items, participants learned 12 items (4 items from each criterion level) in each of three separate blocks of practice. For long-lag items, participants learned all items in one block of practice. After all items in one lag condition had been learned to criterion, participants had initial study and test–restudy practice with items from the second lag condition. Order of presentation of short-lag and long-lag items was counterbalanced across participants.

After all items had been learned to criterion, all participants made *postacquisition aggregate judgments* for each level of lag and criterion level. Prior to making judgments, all participants read detailed instructions regarding lag and criterion level manipulations (see the Appendix). After making aggregate judgments, participants were dismissed and reminded to return 1 week later for the final test. The final test was a participant-paced cued recall test. After the final test, all participants made *posttest aggregate judgments*. As with previous aggregate judgments, participants received instructions regarding lag and criterion level prior to making judgments (see the Appendix).

Results and discussion

No significant differences emerged in any measure as a function of aggregate judgment group (preacquisition vs. no preacquisition), so we collapsed across this variable for all further analyses.

Final test performance The mean percentage of items correctly recalled on the final test as a function of criterion level and lag is presented in Fig. 4. Results of a 2 (lag) × 3 (criterion level) repeated measures ANOVA showed significant main effects of criterion level and lag, as well as a significant interaction, $F(2, 132) = 23.62$, $MSE = .01$, $p < .001$, $F(1, 66) = 90.98$, $MSE = .02$, $p < .001$, and $F(2, 132) = 6.56$, $MSE = .01$, $p = .002$, respectively. Once again, final test performance significantly increased as the number of correct retrievals during practice increased and was significantly higher in the long-lag versus short-lag condition. The interaction again showed that the difference between lag groups was greater as criterion level increased.

Judgments of learning As in Experiment 1, we evaluated the extent to which JOLs are sensitive to the effects of lag and criterion level on final test performance (see Fig. 5). Results of a 2 (lag) × 3 (criterion level) repeated measures ANOVA showed significant main effects of criterion level and lag, as well as a significant interaction, $F(2, 132) = 148.49$, $MSE = 257.16$, $p < .001$, $F(1, 66) = 10.81$, $MSE = 412.96$, $p = .002$, and $F(2, 132) = 3.78$, $MSE = 57.34$, $p = .025$, respectively. Concerning criterion level, as the number of correct retrievals during practice increased, mean JOL values increased, replicating results from Experiment 1. Concerning lag, JOLs were significantly higher for short-lag versus long-lag items. Thus, JOLs showed incorrect directional sensitivity to the effects of lag on final test performance.

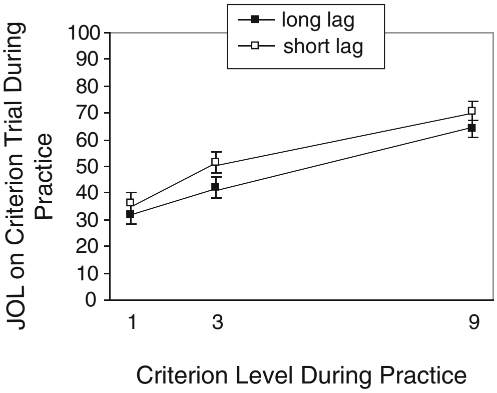

Fig. 5 Mean JOL values on criterion trial during practice as a function of criterion level and lag condition, Experiment 2. Error bars represent standard errors

Concerning the correct directional sensitivity of JOLs to the effects of criterion level on final test performance, to what extent might this relationship reflect metacognitive beliefs about the benefits of more versus fewer correct retrievals during practice? For preacquisition aggregate judgments (leftmost bars in Fig. 6), a repeated measures ANOVA revealed a significant main effect of criterion level, $F(2, 70) = 37.64$, $MSE = 11.06$, $p < .001$. Aggregate judgments increased as a function of criterion level, indicating that participants did have accurate prior metacognitive beliefs about the effects of criterion level on final test performance. The same pattern obtained for postacquisition and posttest aggregate judgments (middle and rightmost bars in Fig. 6), $F(2, 130) = 62.33$, $MSE = 11.07$, $p < .001$, and $F(2, 106) = 45.16$, $MSE = 3.83$, $p < .001$, respectively. Note that although these outcomes establish that participants have appropriate metacognitive beliefs about the effect of criterion level on performance, they do not establish that this knowledge about criterion level

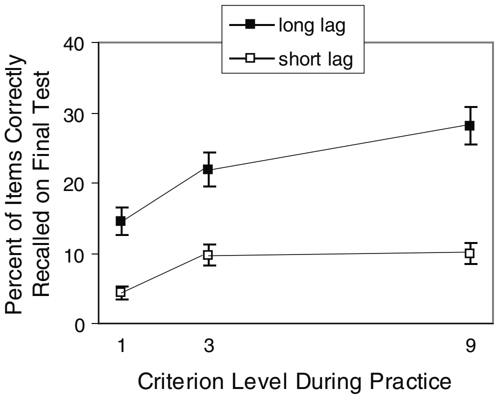

Fig. 4 Mean percentage of items correctly recalled on the final test as a function of criterion level and lag condition, Experiment 2. Error bars represent standard errors

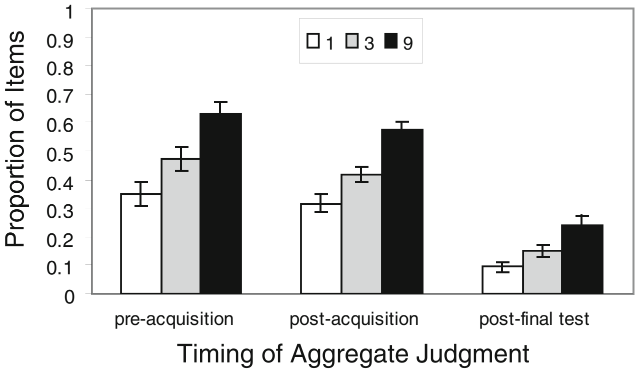

Fig. 6 Mean proportion of criterion level 1, 3, and 9 items that participants predicted they would recall as a function of timing of aggregate judgment, Experiment 2. Error bars represent standard errors

influenced JOLs. However, they do confirm an important precondition by establishing that this extrinsic cue is available for use, which provides additional evidence converging with our interpretation of the outcomes of the HLM analyses (reported below).

Concerning the incorrect directional sensitivity of JOLs to effects of lag on final test performance, participants may have had correct beliefs about lag effects that were overridden by the salient cue of retrieval fluency. Alternatively, participants may have had incorrect or no beliefs about lag effects. Examination of the pattern of aggregate judgments reported in Fig. 7 supports the latter possibility. For preacquisition aggregate judgments (leftmost bars in Fig. 7), no significant differences emerged for short-lag versus long-lag items, $F < 1$. In fact, judgments were almost identical, indicating that participants have no prior beliefs about the memorial benefits of longer versus shorter lags. Postacquisition aggregate judgments also did not differ for long versus short lags (middle bars in Fig. 7), $F < 1$. Of course, at this point, participants have not experienced the memorial benefits of using a longer lag to learn items. After the final test, however, aggregate judgments were still similar for the two lag conditions (rightmost bars in Fig. 7), $F < 1$. Note that although posttest judgments were similar for lag conditions, comparison of posttest judgments with earlier aggregate judgments indicates that participants were learning from experience. Posttest aggregate judgments were lower than preacquisition and postacquisition judgments (ts > 4.82, ps < .001). Thus, the absence of a lag effect was unlikely to have been due to participants' perseverating on prior judgment magnitudes.

JOLs showed correct directional sensitivity to criterion level but showed incorrect directional sensitivity to lag. To what extent were these patterns due to an influence of the mnemonic cue of retrieval fluency on JOLs? To measure retrieval fluency, we examined first keypress latency for all correct retrieval trials in session 1, as in Experiment 1. Figure 8 shows mean first keypress latency (in seconds) as a function of the *n*th correct retrieval during practice. Results of a 2 (lag) × 9 (*n*th correct retrieval) repeated measures ANOVA revealed a significant main effect of lag, $F(1,\ 66) = 125.64$, $MSE = .17$, $p < .001$. First keypress latencies were significantly shorter for the short-lag versus long-lag condition. Results also revealed a significant main effect of the *n*th correct retrieval, as well as a significant interaction, $F(8,\ 528) = 377.71$, $MSE = .07$, $p < .001$, and $F(8,\ 528) = 30.58$, $MSE = .04$, $p < .001$. As in Experiment 1, first keypress latencies significantly decreased as the number of correct retrievals during practice increased.

To examine the extent to which lag, criterion level, retrieval fluency, normative item difficulty, and number of retrieval failures before the first correct recall during practice influenced JOLs, we again conducted HLM analyses. The first model assessed the relationship between lag and JOLs. Consistent with results reported above, the relationship between lag and JOLs was not significant, $t(4744) = 0.91$, $p = .363$. The second model assessed the relationship between criterion level and JOLs. As in Experiment 1, the relationship between criterion level and JOLs was significant, with JOLs increasing as criterion level increased, $t(4744) = 10.63$, $p < .001$. The third model assessed the relationship between first keypress latency and JOLs. As in Experiment 1, the relationship between first keypress latency and JOLs was significant, with JOLs increasing as first keypress latency decreased, $t(4744) = 9.38$, $p < .001$. The fourth model assessed the relationship between normative item difficulty and JOLs and showed a significant relationship,

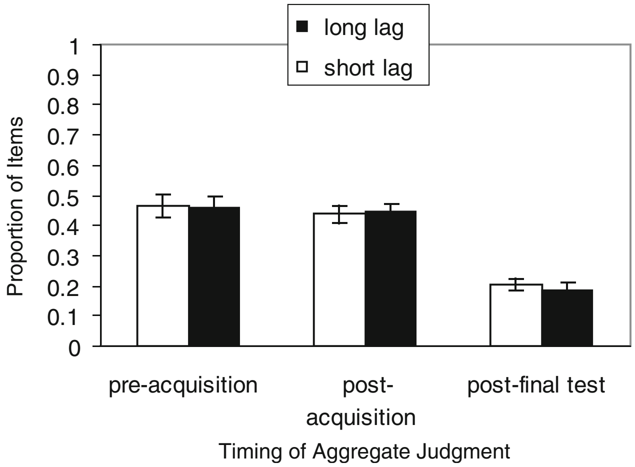

Fig. 7 Mean proportion of short-lag and long-lag items that participants predicted they would recall as a function of timing of aggregate judgment, Experiment 2. Error bars represent standard errors

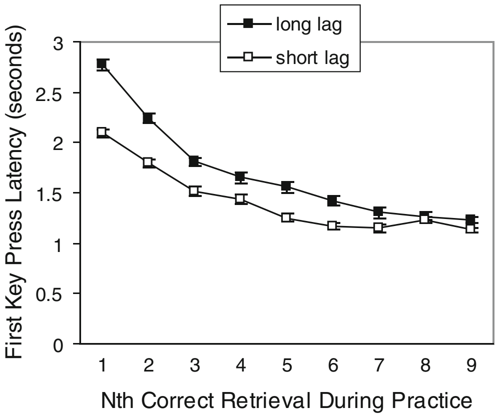

Fig. 8 Mean first keypress latency (in seconds) as a function of the *n*th correct retrieval during practice for each lag condition, Experiment 2. Error bars represent standard errors

with higher JOLs for normatively easier versus more difficult items, $t(4744) = 2.67$, $p = .004$. The fifth model assessed the relationship between number of incorrect retrievals during encoding and JOLs and showed a nonsignificant relationship, $p > .05$.

Of greatest interest, we ran a sixth model with all variables that were significantly related to JOLs to assess the extent to which each variable uniquely influenced JOLs. Results showed that criterion level, first keypress latency, and normative item difficulty were all significantly related to JOLs, $t(4742) = 26.16$, $p < .001$, $t(4742) = 6.63$, $p < .001$, and $t(4742) = 2.43$, $p = .015$, respectively. These results suggest that the extrinsic cues of criterion level, the intrinsic cue of normative item difficulty, and the mnemonic cue of retrieval fluency each uniquely influenced JOLs. Why did normative item difficulty influence JOLs in the present experiment, but not in Experiment 1? Because lag was manipulated within participants, Experiment 2 included more items than did Experiment 1, which led to the inclusion of more difficult items. Results from the HLM suggest that having a larger range of item difficulty may have provided participants with another cue for making JOLs. Most important, however, we replicated the results from Experiment 1, with both criterion level and first keypress latency influencing JOLs.

General discussion

The present experiments evaluated two questions. First, are JOLs made after correct retrievals during practice sensitive to the effects of quantity and timing of these correct retrievals on final test performance? Second, which cues are used to make JOLs during criterion learning? Concerning the quantity of correct retrievals, JOLs showed correct directional sensitivity to the effects of criterion level on final test performance: Both performance and JOLs increased as the number of correct retrievals during practice increased. In contrast, concerning the timing of correct retrievals, JOLs did not show correct directional sensitivity to the effects of lag on final test performance: Performance was greater for items correctly retrieved after longer versus shorter lags, but JOLs were not, with numerical trends (Experiment 1) or significant differences (Experiment 2) in the opposite direction. In relation to the second question, results from both Experiments 1 and 2 showed that the mnemonic cue of retrieval fluency and the extrinsic cue of criterion level influenced JOLs. Additionally, the intrinsic cue of item difficulty influenced JOLs in Experiment 2, when items had a wider range of difficulty. However, the extrinsic cue of lag did not influence JOLs, nor did the cue of number of failed retrieval attempts during practice.

Given that both JOLs and aggregate judgments showed correct directional sensitivity to the effects of criterion level on final test performance, additional research exploring the extent to which individuals use their metacognitive knowledge about the memorial benefits of increasing criterion levels to control self-regulated retrieval practice will be informative. For example, Kornell and Bjork (2008) had participants learn items for a later retention test and allowed some participants to drop items from further practice during learning. Results showed that a majority of items were dropped from practice after one correct recall. This result is somewhat troubling, given the substantial gains in final test performance after an item has been correctly recalled more than one time during practice.

Why did Kornell and Bjork's (2008) participants drop items after only one correct recall during practice when participants in the present study demonstrated metacognitive knowledge about the memorial benefits of more versus fewer correct retrievals during learning? One possibility is that Kornell and Bjork's participants were being strategic on the basis of the time constraints imposed in that study. Specifically, participants were given only 10 min to learn as many items as they could. Participants may have discontinued practice with items after they could correctly recall them one time so that they could focus the remainder of their limited study time on items that had not yet been correctly recalled. If participants were given the goal of learning items for a later retention test and also were given unlimited time to learn the items, it is possible that their self-regulated decisions would more closely resemble their judgments for criterion level in the present experiments, with participants deciding to practice items until they are correctly recalled multiple times before dropping them from practice. Nonetheless, these results leave open the possibility that participants may not effectively self-regulate practice, even though results from the present experiments demonstrate that participants have metacognitive knowledge regarding criterion level effects.

Although the results reported here consistently demonstrated that metacognitive judgments showed correct directional sensitivity to the effects of criterion level on later performance, the results from Karpicke (2009) showed a different pattern. To revisit, after items were learned to criterion during practice, participants were asked to make aggregate judgments. Results showed that judgments did not differ for individuals who terminated practice after one correct recall versus those who completed two additional practice trials. What might explain the inconsistency between the present findings and those of Karpicke? One possibility is that differences in experimental design influenced metacognitive judgments. Criterion level was a within-participants manipulation in the present experiments, whereas practice schedule was a between-participants manipulation in Karpicke's study. Previous research has shown that the extent to which individuals incorporate metacognitive beliefs into their metacognitive judgments can depend on the extent to which the encoding conditions elicit attention to a given variable (e.g., Koriat et al., 2004). Because within-participants manipulations allow participants to experience different levels of a variable (e.g., criterion level), they may be more likely to consider their beliefs about the effects of that variable when making judgments than in a between-participants design. Although additional research will be needed to explore this possibility further, this account does provide a plausible reconciliation of the apparent inconsistency between the present outcomes and those of Karpicke.

Not only would future research evaluating the sensitivity of JOLs to the effects of criterion level on final test performance be beneficial for understanding self-regulated decisions individuals make when they have unlimited time to learn items, but also it would provide insight into the pattern of diminishing returns observed for criterion level. Previous research has shown that performance increases as criterion level increases, but the incremental benefit to final test performance decreases as criterion level increases (i.e., Figs. 1 and 4; see also Pyc & Rawson, 2009; Vaughn & Rawson, 2011). Although JOLs in the present experiments showed correct directional sensitivity to the effects of criterion level on final test performance in that they increased as criterion level increased, they did not appropriately reflect the diminishing returns of increasing criterion level on final test performance (e.g., the curvilinear pattern of final test performance in Fig. 4 vs. the linear pattern of JOLs in Fig. 5). What does this suggest about the potential basis for JOLs, given that they did not properly reflect the pattern of diminishing returns for final test performance? On one hand, one might think that the pattern reflects the influence of ease of processing during retrieval (given that latencies decrease as criterion level increases). However, the finding that criterion level still influenced JOLs even after controlling for retrieval latency in the HLM analyses weighs against this account. Another possibility suggested by the aggregate judgments is that students have incorrect beliefs about this particular feature of criterion level effects. Future research could evaluate why JOL are not sensitive to the diminishing returns of increasing criterion level.

In the presenst experiments, JOLs did not show correct directional sensitivity to the effects of lag on final test performance. This lack of sensitivity to lag is unfortunate, given the substantial effects of lag on performance. Note that final test performance following just one correct recall at a long lag was as good as (Experiment 1) or even better than (Experiment 2) nine correct recalls at a short lag. Students are likely not fully capitalizing on the benefits of testing by not appreciating the influence of lag on retention. This possibility is further bolstered by complementary results reported by Kornell (2009). Across a series of experiments, Kornell evaluated the effectiveness of a fixed number of practice trials administered with either a short or a long lag. Of interest here, participants made aggregate judgments after studying items during practice. Results showed that although final test performance was greater for items practiced with longer versus shorter lags, aggregate judgments were higher for items practiced with shorter versus longer lags. Taken together, these results demonstrate that participants do not understand the memorial benefits of longer versus shorter lags during learning. These results are somewhat troubling, given that the lag effect is one of the most robust findings in the memory literature and has the potential to greatly impact student learning and scholarship.

Although JOLs were of primary interest in the present experiments, in Experiment 2 we also included aggregate judgments in order to more directly evaluate beliefs about criterion level and lag. These aggregate judgments may be useful for further investigating beliefs about other factors in future research. For example, aggregate judgments may shed light on the extent to which individuals are aware of the memorial benefits of difficult retrievals during encoding. In the present experiments, retrieval fluency consistently had an influence on judgments (i.e., judgments increased as fluency decreased). In keeping with Koriat's (1997) cue-utilization framework, we classified retrieval fluency as a mnemonic cue. However, some research has shown that the relationship between fluency and JOLs may be theory driven (i.e., based on beliefs; e.g., Matvey, Dunlosky, & Guttentag, 2001). Future research could evaluate what people believe about the effect of fluency on performance by including aggregate judgments about retrieval fluency.

The present experiments extend beyond prior metacognitive research by evaluating the sensitivity of JOLs to the effects of the quantity and timing of successful retrievals during criterion learning (as opposed to prior research involving study only and/or fixed amounts of practice trials) on final test performance. The present results indicate that JOLs made after correct retrievals during practice show correct directional sensitivity to the effects of criterion level on final test performance but do not show correct directional sensitivity to the effects of lag on final test performance. Given the important implications for student learning and scholarship, one goal of future research should be to evaluate ways to improve the sensitivity of JOLs to the effects of lag on final test performance. Additionally, future retrieval practice research should evaluate the extent to which individuals' self-regulated decisions are related to their JOLs and beliefs.

Appendix. Instructions for participants in Experiment 2

Description of criterion level and lag

During this experiment you will be asked to practice items until you have correctly recalled them multiple times (1, 3, or 9 times).

For some of the word pairs, there will be 11 other items between each next practice trial with a given word pair. For other word pairs, there will be 35 other items between each practice trial with a given word pair.

Another way that you can think about the number of word pairs between each practice trial is by pretending you are using a deck of flashcards to study. Imagine you had a deck of 12 flashcards with one word pair on each card. When you finish practicing with the first card, you place it at the bottom of the deck. After you practice the other 11 cards in the deck, the first card is back at the top of the stack and you study it again. Alternatively, imagine you had a deck of 36 flashcards. In this case, you would end up studying 35 cards after the first one before you got around to it again.

Lag aggregate judgment screen for preacquisition judgment

For 36 of the word pairs you practice today, there will be 35 items between each next practice trial with a given item (like practicing with one deck of flashcards that has 36 cards in it). How many of these 36 items will you be able to remember on the test you will take **one week from today**? Enter answer in box provided below: (Remember, value must be between 0 and 36)

For 36 of the word pairs you practice today, there will be 11 items between each next practice trial with a given item (like practicing with three decks of flashcards that each have 12 cards in them). How many of these 36 items will you be able to remember on the test you will take **one week from today**? Enter answer in box provided below (Remember, value must be between 0 and 36).

Criterion level aggregate judgment screen for preacquisition judgment

During this experiment, you will be asked to learn items until they are correctly recalled 1, 3, or 9 times during practice. Please read each question below carefully and indicate how many of the items that are recalled 1, 3, or 9 times during practice you think you will be able to recall on a test one week from today.

1) How many items (out of 24) that are correctly recalled **1** time during practice do you think you will be able to correctly recall on a test **one week from today**? Type answer in box provided below.
2) How many items (out of 24) that are correctly recalled **3** times during practice do you think you will be able to correctly recall on a test **one week from today**? Type answer in box provided below.
3) How many items (out of 24) that are correctly recalled **9** times during practice do you think you will be able to recall on a test **one week from today**? Type answer in box provided below.

Note For postacquisition and posttest aggregate judgments, participants saw similar instructions, with the exception that all statements indicating that they would be learning items were changed to the past tense (i.e., for items that you learned during practice or for items that you learned last week). For the postfinal test aggregate judgments, we asked participants how many items they believed that they had recalled, as opposed to how many items they thought that they would remember in 1 week.

Author Notes The research reported here was supported by a James S. McDonnell Foundation 21st Century Science Initiative in Bridging Brain, Mind and Behavior Collaborative Award, and by the Institute of Education Sciences, U.S. Department of Education, through Grant R305H050038 to Kent State University. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

Thanks to Nicole Gonzalez, Phil Grimaldi, Rochelle O'Neil, Melissa Roderick, and Sara Smith for assistance with data collection. Thanks also to Jeff Ciesla for assistance with analyses. A special thanks to John Dunlosky and RADlab for helpful comments on earlier versions of the manuscript.

References

Allen, G. A., Mahler, W. A., & Estes, W. K. (1969). Effects of recall tests on long-term retention of paired associates. *Journal of Verbal Learning and Verbal Behavior, 8,* 463–470.

Ariel, R., Dunlosky, J., & Bailey, H. (2009). Agenda-based regulation of study-time allocation: When agendas override item-based monitoring. *Journal of Experimental Psychology. General, 138,* 432–447.

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.

Cepeda, N. J., Pashler, H., Vul, E., Wixted, J. T., & Rohrer, D. (2006). Distributed practice in verbal recall tasks: A review and quantitative synthesis. *Psychological Bulletin, 132,* 354–380.

Cull, W. (2000). Untangling the benefits of multiple study opportunities and repeated testing for cued recall. *Applied Cognitive Psychology, 14,* 215–235.

Dunlosky, J., & Metcalfe, J. (2009). *Metacognition*. Beverly Hills, CA: Sage.

Dunlosky, J., & Rawson, K. A. (in press). Overconfidence produces underachievement: Inaccurate self evaluations undermine students' learning and retention. *Learning and Instruction*

Greene, J. A., & Azevedo, R. (2007). A theoretical review of Winne and Hadwin's model of self-regulated learning: New perspectives and direction. *Review of Educational Research, 77,* 334–372.

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., & Roediger, H. L., III. (2007). Repeated retrieval during learning is the key to long-term retention. *Journal of Memory and Language, 57,* 151–162.

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

Koriat, A. (1997). Monitoring one's own knowledge during study: A cue-utilization approach to judgments of learning. *Journal of Experimental Psychology. General, 126,* 349–370.

Koriat, A., Bjork, R. A., Sheffer, L., & Bar, S. K. (2004). Prediction one's own forgetting: The role of experience-based and theory-based processes. *Journal of Experimental Psychology. General, 133,* 643–656.

Kornell, N. (2009). Optimizing learning using flashcards: Spacing is more effective than cramming. *Applied Cognitive Psychology, 23,* 1297–1317.

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

Kornell, N., Hays, M. J., & Bjork, R. A. (2009). Unsuccessful retrieval attempts enhance subsequent learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35,* 989–998.

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

Landauer, T. K., & Bjork, R. A. (1978). Optimum rehearsal patterns and name learning. In M. M. Gruneberg, P. E. Morris, & R. N. Sykes (Eds.), *Practical aspects of memory* (pp. 625–632). New York: Academic Press

Leonesio, R. J., & Nelson, T. O. (1990). Do different metamemory judgments tap the same underlying aspects of memory? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16,* 464–470.

Matvey, G., Dunlosky, J., & Guttentag, R. (2001). Fluency of retrieval at study affects judgments of learning (JOLs): An analytic or nonanalytic basis for JOLs? *Memory & Cognition, 29,* 222–233.

Mazzoni, G., Cornoldi, C., & Marchitelli, G. (1990). Do memorability ratings affect study-time allocation? *Memory & Cognition, 18,* 196–204.

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.

Meeter, M., & Nelson, T. O. (2003). Multiple study trials and judgments of learning. *Acta Psychologica, 113,* 123–132.

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

Nelson, T. O., & Dunlosky, J. (1994). Norms of paired-associate recall during multitrial learning of Swahili–English translation equivalents. *Memory, 2,* 325–335.

Nelson, T. O., Leonesio, R. J., Shimamura, A. P., Landwehr, R. F., & Narens, L. (1982). Overlearning and the feeling of knowing. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 8,* 279–288.

Nelson, T. O., & Narens, L. (1990). Metamemory: A theoretical framework and new findings. *Psychology of Learning and Motivation, 26,* 125–172.

Pashler, H., Zarrow, G., & Triplett, B. (2003). Is temporal spacing of test helpful even when it inflates error rates? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29,* 1051–1057.

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

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.

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

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.

Rawson, K. A., Dunlosky, J., & McDonald, S. L. (2002). Influences of metamemory on performance predictions for text. *Quarterly Journal of Experimental Psychology, 55A,* 505–524.

Rawson, K. A., O'Neil, R. L., & Dunlosky, J. (2011). Accurate monitoring leads to effective control and greater learning of patient education materials. *Journal of Experimental Psychology. Applied, 17,* 288–302.

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

Schwartz, J. E., & Stone, A. A. (1998). Strategies for analyzing ecological momentary assessment data. *Health Psychology, 17,* 6–16.

Shaughnessy, J. J., & Zechmeister, E. B. (1992). Memory-monitoring accuracy as influenced by the distribution of retrieval practice. *Bulletin of the Psychonomic Society, 30,* 125–128.

Thiede, K. W. (1999). The importance of monitoring and self-regulation during multitrial learning. *Psychonomic Bulletin & Review, 6,* 662–667.

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

Thiede, K. W., & Dunlosky, J. (1999). Toward a general model of self-regulated study: An analysis of selection of items for study and self-paced study time. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24,* 1024–1037.

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.

Wheeler, M. A., & Roediger, H. L., III. (1992). Disparate effects of repeated testing: Reconciling Ballard's (1913) and Bartlett's (1932) results. *Psychological Science, 3,* 240–245.

Winne, P. H., & Hadwin, A. F. (1998). Studying as self-regulated learning. In D. J. Hacker, J. Dunlosky, & A. Graesser (Eds.), *Metacognition in educational theory and practice* (pp. 277–304). Mahwah, NJ: Erlbaum.

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