Measuring memory monitoring with judgements of retention (JORs)

Sarah K. Tauber1 and Matthew G. Rhodes2

1Department of Psychology, Kent State University, Kent, OH, USA
2Department of Psychology, Colorado State University, Fort Collins, CO, USA

Most prior research has examined predictions of future memory performance by eliciting judgements of learning (JOLs). In six experiments, we explored monitoring with an alternative prospective measure. Specifically, participants made judgements of retention (JORs) predicting how long (in min) they would be able to remember information. Results revealed that participants provided relatively short predictions of how long they would remember information. Further, participants’ JORs were sensitive to recall performance as well as manipulations that influenced memory performance indicating that they were able to effectively monitor learning using JORs. JORs influenced study decisions as well, with participants selecting more items for restudy following JORs than following JOLs or no monitoring judgement. However, restudy selection did not vary between a JOR and a JOL condition framed in terms of forgetting. Thus, we suggest that, much like forget-framed JOLs, JORs may bring different information—such as memory failure—to mind. In all, the inferential mechanisms underlying metacognitive monitoring with JOLs extends to monitoring when measured with JORs. Assessing monitoring with JORs provides information not available with JOLs (i.e., memory duration estimates) and a different basis for study decisions from remember-framed JOLs.

Keywords: Metacognition; Metamemory; Monitoring; Judgements of retention; Retention; Forgetting; Self-regulation.

Metamemory refers to awareness of one’s own memory processes, which includes monitoring and control (Dunlosky & Metcalfe, 2009; Koriat, 2007; Metcalfe, 2000; Nelson & Narens, 1994). Monitoring plays a vital role in promoting effective study choices and, thus, influences memory performance. For example, a student will be better prepared for an upcoming exam if he or she is able to accurately assess (i.e., monitor) learning in order to make optimal study decisions and perform well on the exam. Unfortunately for our example student, prior research has demonstrated
that, under some circumstances, participants’ assessments of learning are insensitive to critical factors that influence memory performance (Carroll, Nelson, & Kirwan, 1997; Koriat, Bjork, Sheffer, & Bar, 2004; Shadduck & Carroll, 1997). Such conclusions about monitoring have been based almost exclusively on predictions of future memory performance commonly made on a percentage scale, elicited using judgements of learning (JOLs). In the experiments reported, we assessed the underlying bases of participants’ self-regulated learning with a new measure of monitoring (i.e., a judgement of retention, JOR), instructing participants to indicate how long information would be remembered. We first provide an overview of monitoring and control, then consider prior work examining the degree to which monitoring is influenced by information about forgetting, and finally review alternative measures to JOLs.

**Monitoring and control**

Monitoring processes are commonly assessed using JOLs. In a typical experiment, a participant makes a JOL by predicting how likely he or she is to later remember an item (e.g., I am 80% likely to remember this word). Monitoring (as assessed via JOLs) is often accurate, yet a number of discrepancies between predictions and actual memory performance have been demonstrated (e.g., Carroll et al., 1997; Kornell, Rhodes, Castel, & Tauber, 2011; Rhodes & Castel, 2008, 2009; Tauber & Rhodes, 2010a, 2010b, 2011). For example, JOLs may be based on inferences regarding information available during encoding (cue-utilization framework; Kornell, 2011). That is, in order to monitor their learning, participants may use various cues, such as features of the items to be remembered, as the basis for their JOLs. Thus, errors in monitoring often occur because predictions can be based on a variety of cues that are salient when making the judgement but are unrelated to future memory performance, or because predictions are insensitive to cues that are diagnostic of future memory performance (but see, e.g., Nelson & Dunlosky, 1991; Rhodes & Tauber, 2011).

For example, to evaluate monitoring of future forgetting, Koriat et al. (2004, Experiment 1) presented participants with related (e.g., CAT–KITTEN) and unrelated (e.g., TABLE–MONKEY) word pairs and solicited JOLs based on one of three retention intervals (RIs): immediate recall, recall in 1 day, or recall in 1 week. Memory was tested following the specified RI, and results demonstrated that JOLs and recall were higher for related than for unrelated pairs. With regard to the RI, recall was poorer for longer RIs, whereas participants’ JOLs were entirely insensitive to RI. However, when a within-subjects manipulation was employed, affording participants the opportunity to compare RIs, results indicated that JOLs decreased as the RI increased. That is, JOLs were lowest for a 1-week delay, moderate for a 1-day delay, and highest for a 10-min delay. These data suggest that participants have metacognitive knowledge that a long RI between a study and test can increase forgetting (cf. Rawson, Dunlosky, & McDonald, 2002), but this knowledge only becomes evident when participants can compare RIs of different lengths.

As measured with JOLs, participants’ memory predictions appear to reflect forgetting under limited circumstances. However, such monitoring errors may reflect the measure itself, rather than participants’ awareness of the influence of an intervening retention interval on their memory. In contrast with JOLs, judgements of retention (JORs) may be sensitive to forgetting in a manner that is not manipulation bound. That is, because time is made salient in the judgement itself (i.e., participants indicate how long each word will be remembered), the durations that participants provide reflect appraisals of the durability of memory for each item. Thus, JORs disclose participants’ beliefs about memory duration for each item and provide an assessment of metacognitive awareness of other factors (e.g., item characteristics) that influence later memory performance as well.

To what extent do measures of monitoring influence metacognitive control, such as study choices? Control is often measured via study termination or the selection of items for restudy (Nelson & Narens, 1994), and prior work suggests...
that monitoring influences control processes (e.g., Metcalfe & Finn, 2008; Rhodes & Castel, 2009). In addition, the specific monitoring judgement used can modify control decisions made during study. For example, Finn (2008) asked participants to make judgements of either how likely they were to remember a word (remember JOL) or how likely they were to forget a word (forget JOL; see also Koriat et al., 2004). Following the memory judgement, participants in both conditions selected items for restudy prior to a memory test. Results indicated that forget-framed JOLs more closely corresponded with later recall (i.e., participants were better calibrated) than did remember-framed JOLs (but see Kornell & Bjork, 2009; Rhodes & Castel, 2008). Further, participants chose to restudy more items when making forget JOLs than when making remember JOLs. Finn (2008) suggested that predictions of forgetting made memory failure more salient, reducing the magnitude of predictions of later recall and influencing control choices relative to predictions of remembering (cf. Friedman & Castel, in press).

Similar effects might drive restudy decisions when participants monitor their learning with predictions of memory performance in units of time. In particular, JORs could prime instances of memory failure in a similar manner to forget-framed JOLs. Thus, participants may make different study choices compared with a JOL condition. For example, participants might select more items for restudy after making JORs than after making JOLs. Thus, in addition to assessing monitoring via JORs, a primary goal for the current work was to determine whether JORs differentially impact study choices in comparison with JOLs (Experiments 5 and 6).

**Alternative measures of monitoring**

The current work is not the first to employ alternative measures in an effort to gain information about participants’ monitoring that cannot be obtained from a JOL. A few prior studies have focused on alternatives to JOLs for eliciting predictions of future memory performance (Kelemen, 2000; McCabe & Soderstrom, 2011; McGillivray & Castel, 2011; Tiede & Leboe, 2009) or alternatives to assess monitoring of forgetting (Finn, 2008; Halamish, McGillivray, & Castel, 2011). Tiede and Leboe explored monitoring over multiple study–test opportunities (cf. Koriat, Sheffer, & Ma’ayan, 2002). Specifically, Tiede and Leboe’s participants studied the same list of words, made JOLs, and tested their memory in four study–test trials. Further, participants were asked to judge the rate of change in memory performance (cf. Metcalfe & Kornell, 2005; Townsend & Heit, 2011), predicting how many more or fewer word pairs they expected to recall following an additional study opportunity. Results demonstrated that participants’ judgements of the rate of change in recall better predicted the benefits of future learning than did JOLs. Tiede and Leboe suggested that the framing of the prediction made additional study opportunities salient, encouraging participants to incorporate their beliefs about additional study into their change-in-recall judgements.

Additionally, Halamish et al. (2011) introduced an alternative measure that was specifically intended to measure monitoring of forgetting (cf. Finn, 2008). In particular, young and older adult participants studied items from five categories (e.g., vegetables) in preparation for a free recall test. Following the test participants made a monitoring-of-forgetting (MOF) judgement indicating how many more words from each category they could not remember. Of particular interest, MOF judgements were closely related to the number of items that were forgotten for both age groups even though older adults forgot more items than did young adults. These data suggest that ageing did not impact participants’ ability to assess forgetting and, moreover, that the MOF judgement may be a useful technique to assess posttest monitoring of forgetting.

Judgements of the rate of change and MOF judgements provide information that is not readily available with JOLs alone. However, these measures do not permit an item-by-item examination of monitoring. That is, these judgements are made for aggregate sets of items, but with such global measures, the accuracy of participants’ memory judgements for one item compared with
another (i.e., relative accuracy) cannot be evaluated. This limitation is remedied with the use of JORs that are made on an item-by-item basis. Thus, JORs are a richer measure, relative to other alternative measures, because they provide item-by-item assessments of memory duration. Further, JORs may provide additional information about monitoring of memory duration and monitoring of forgetting that is unattainable with JOLs.

The current research
Considering memory durability during learning could be valuable in a number of instances, such as occasions in which you are introduced to a person that you are likely to see again in a month (but not the next week). Thus, the primary motivation for the current research was to introduce a new, time-based judgement, JOR, and to explore the bases for participants’ judgements with this measure. In the experiments reported, participants studied to-be-remembered material, provided a JOR indicating how long, typically in minutes, an item would be remembered, and then received a memory test. Our primary goal was to examine the information participants used in making JORs. In particular, we assessed whether JORs were inferential and based on the cues present during study (Koriat, 1997). For example, in all six experiments, we assessed whether participants’ JORs were sensitive to factors that influence memory performance. We evaluated this by conditionalizing JORs based on memory performance (cf. Daniels, Toth, & Hertzog, 2009), and via item-by-item gamma correlations (i.e., resolution).

It was expected that participants would be able to monitor their learning with JORs such that JORs would be longer for words that were later remembered than for words that were not later remembered. Further, JORs and recall should be positively related on an item-by-item basis, suggesting that words given longer JORs were more likely to be remembered than words given shorter JORs. In Experiments 5 and 6, we evaluated metacognitive control (i.e., restudy choice) in three judgement conditions: JORs, JOLs, and a condition with no monitoring judgement. We further investigated metacognitive control processes in Experiment 6 by comparing study decisions following JORs, JOLs, and JOLs framed in terms of forgetting. It was expected that study decisions following JORs would be different from study decisions following JOLs or with no monitoring judgement (restudy only condition). Specifically, we expected JORs to bring different information to mind, with the consequence that participants might select more items for restudy following JORs than following JOLs or no monitoring judgement. Taken together, the experiments reported provide an in-depth examination of monitoring accuracy, introduce and evaluate a new measure of monitoring (i.e., JORs), and assess the influence of metacognitive judgements on metacognitive control processes.

EXPERIMENT 1
Participants in Experiment 1 studied a list of words that they were to remember for an upcoming test. For each word, participants provided a JOR, indicating how long they would remember the item. A minute scale (0–60 min) was used for JORs.

Method
Participants
Forty Colorado State University students (\(M_{\text{age}} = 19.45\) years, \(SD = 1.30\)) received credit in psychology courses for participation.

Materials and procedure
Participants were presented with 30 nouns from one of two randomized versions of the study list (word length: \(M = 4.77, SD = 1.07\); word frequency: \(M = 37.54, SD = 12.24\); Kucera & Francis, 1967). No significant differences were found in recall based on item order (\(t < 1\)) for Experiment 1 or for any other experiment reported and are not discussed further. Participants made JORs on a minute scale (0–60 min), where a judgement of 0 indicated not being able to remember the item immediately at all, and a judgement of 60
indicated the ability to remember an item in 60 min (1 hr). The first two and the last two words of the study list were treated as buffers and were excluded from all analyses reported. Words were presented one at a time for 4 s; participants were then given 5 s to write down their JOR. A 500-ms interstimulus interval (ISI) preceded the presentation of the next study item after the JOR. Participants received a 3-min filler task (listing states of the United States), following which participants were allotted 3 min for free recall.

Results and discussion

JORs and recall

The alpha level was set at .05 for this and all subsequent experiments. Participants provided relatively short JORs (M = 15.12 min, SD = 11.79 min), given the range of the JOR scale (0–60 min), and the JORs were reliably greater than zero, t(39) = 8.11 (see Figure 1a). JORs were conditionalized based on recall to create an average JOR for items that were subsequently remembered and an average JOR for items that were subsequently forgotten (see Table 1). Average JORs were reliably longer for items that were remembered than for items that were not remembered, t(32) = 5.48, Cohen’s d = 0.52. Finally, participants were permitted to predict how long they would remember words, the durations specified were relatively modest. Moreover, participants’ JORs were longer for items that were recalled than for items that were not recalled.

Resolution

Resolution (relative accuracy) is typically measured via within-subjects Goodman–Kruskal gamma correlation (Nelson, 1984; but see Benjamin & Diaz, 2008; Masson & Rotello, 2009, for alternatives). Measures of resolution reflect how well people predict performance on an item-by-item level, capturing the degree to which predictions distinguish between items that will and will not be remembered. A one-sample t test indicated that average gamma correlations (see Table 2) were significantly greater than zero, t(39) = 5.28, p < .001. Thus, JORs and recall were reliably related as participants were significantly better than chance at determining whether they would be able to remember each word.

EXPERIMENT 2

Experiment 1 demonstrated that participants successfully monitored their learning with JORs. Specifically, JORs were longer for words that were remembered than for words that were not remembered. Further, JORs were positively related to recall performance on an item-by-item basis. However, participants in Experiment 1 were constrained to a 0–60-min scale for their judgements. With an unconstrained scale, participants’ JORs may be unrelated to memory performance. Thus, in Experiment 2 we used an identical procedure to that in Experiment 1, with the exception that participants were allowed to use any time scale (i.e., seconds, minutes, hours, etc.) for JORs.

Method

Participants

Thirty-two Colorado State University students (M age = 18.56 years, SD = 0.72) received credit in psychology courses for participation.

Materials and procedure

The materials and procedure were identical to those in Experiment 1 except for the scale used for JORs. In particular, participants in Experiment 2 made
JORs using any time scale desired for each item to be remembered. Thus, a participant could indicate that he or she would remember one word for 1 hr, another word for 20 min, and yet another word for 30 s, and so on. Participants were required to identify the time scale used (e.g., minutes, hours) for each judgement, and all time estimates were converted to a minute scale in order to compare JORs across participants.

Results and discussion

JORs and recall

Participants generally provided short JORs ($M = 6.96$ min, $SD = 20.65$ min) that were reliably greater than zero, $t(33) = 3.58$ (see Figure 1b). The majority of participants (84.4%) used seconds or minutes for their JORs (34.4% always used a minute scale, 9% always used a seconds scale, and...
Of the remaining participants, 6.25% used a combination of minutes and hours, 6.25% used a combination of seconds, minutes, and hours, and finally 3.1% (1 participant) consistently used hours. Similar to prior experiments, JORs were conditionalized based on recall performance (see Table 1). Average JORs were reliably longer for items that were remembered than for items that were not remembered, $t(31) = 3.12$, Cohen’s $d = 0.11$. Finally, participants recalled approximately 27% ($SD = 14.08\%$) of the items studied.

**Table 1. Average JORs conditionalized by recall performance**

<table>
<thead>
<tr>
<th>Experiment</th>
<th>Recalled</th>
<th>Unrecalled</th>
</tr>
</thead>
<tbody>
<tr>
<td>Experiment 1</td>
<td>15.78 (1.79)</td>
<td>10.99 (1.41)</td>
</tr>
<tr>
<td>Experiment 2</td>
<td>9.16 (4.11)</td>
<td>6.70 (3.64)</td>
</tr>
<tr>
<td>Experiment 3</td>
<td>23.70 (2.81)</td>
<td>15.16 (1.40)</td>
</tr>
<tr>
<td>Experiment 4</td>
<td>15.65 (1.79)</td>
<td>11.18 (1.58)</td>
</tr>
<tr>
<td>Abstract items</td>
<td>21.12 (2.40)</td>
<td>15.73 (1.74)</td>
</tr>
<tr>
<td>Concrete items</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Note: JORs = judgements of retention, in minutes. Standard errors provided in parentheses. In all experiments, reliably higher JORs were provided for recalled words than for unrecalled words.*

41% used a combination of minutes and seconds). Of the remaining participants, 6.25% used a combination of minutes and hours, 6.25% used a combination of seconds, minutes, and hours, and finally 3.1% (1 participant) consistently used hours.

Similar to prior experiments, JORs were conditionalized based on recall performance (see Table 1). Average JORs were reliably longer for items that were remembered than for items that were not remembered, $t(31) = 3.12$, Cohen’s $d = 0.11$. Finally, participants recalled approximately 27% ($SD = 14.08\%$) of the items studied.

A one-sample $t$ test indicated that average gamma correlations (see Table 2) were significantly greater than zero, $t(31) = 4.43$, $p < .001$. Thus, participants were significantly better than chance at determining whether they would be able to retain each word.

**EXPERIMENT 3**

Replicating Experiment 1, Experiment 2 demonstrated that participants effectively monitored their learning with JORs. That is, in both experiments, participants’ JORs discriminated between items that were remembered and items that were not remembered. Further, participants in Experiments 1 and 2 did not provide extreme predictions of memory duration (cf. Carroll et al., 1997; Koriat et al., 2004; Shaddock & Carroll, 1997). Indeed, when participants were allowed to provide any measure of time as the basis for their JORs, they generally made even shorter predictions (i.e., minutes, seconds) than those evident in Experiment 1.

In Experiments 1 and 2, participants received a 3-min retention interval (RI). However, JORs are time-based predictions, and it is possible that the anticipated RI may influence these judgements. Thus, in Experiment 3, we explored whether the accurate monitoring from Experiments 1 and 2 would generalize to a longer RI (i.e., 20-min RI). Further, in contrast to Experiments 1 and 2, participants in Experiment 3 were provided with information about the retention interval. Thus, in Experiment 3 we evaluated whether participants could monitor their learning with JORs when the anticipated RI was made clear.

**Method**

**Participants**

Twenty-six Colorado State University students ($M_{age} = 19.00$ years, $SD = 1.44$) received credit in psychology courses for participation.

**Materials and procedure**

The materials used in Experiment 3 were identical to those in Experiments 1 and 2. Due to the increased variability in JORs when no scale was
provided (Experiment 2), the remaining experiments used the 0–60-min scale employed in Experiment 1. Participants in Experiment 3 anticipated and received a 20-min retention interval that was filled with a series of four filler tasks (5 min completing maths problems, 5 min listing states of the United States, 5 min completing a novel set of math problems, and 5 min listing major cities of the United States). All other aspects of the procedure were identical to those in Experiment 1.

Results and discussion

JORs and recall

Given that the JOR scale extended to 60 minutes, participants provided relatively short JORs ($M = 18.18$ min, $SD = 10.46$ min) that were reliably greater than zero, $t(25) = 8.86$ (see Figure 1c). Analyses with conditionalized JORs based on whether each item was recalled or not (see Table 1) indicated that average JORs were reliably longer for items that were remembered than for items that were not remembered, $t(24) = 4.82$, Cohen’s $d = 0.77$.

Participants recalled approximately 22% of the items studied ($M = 22.17$, $SD = 8.43$). Recall performance in Experiment 3 was reliably lower than the combined recall performance in Experiments 1 and 2 in which a 3-min RI was employed, $t(96) = 3.06$, Cohen’s $d = 0.70$. Thus, our manipulation of a longer retention interval was effective at reducing the overall level of memory performance.

Resolution

A one-sample $t$ test showed that average gamma correlations were significantly greater than zero, $t(25) = 7.21$, $p < .001$ (see Table 2), indicating that participants were significantly better than chance at determining whether they would be able to retain each word.

EXPERIMENT 4

Experiment 3 demonstrated that participants effectively monitored their learning with JORs, even when RI information was provided prior to study, with a longer RI, and with lower overall recall performance. Specifically, replicating prior experiments, JORs were reliably longer for words that were remembered than for words that were not remembered. Additionally, JORs and recall were related on an item-by-item basis with results indicating that words given longer JORs were more likely to be remembered than words given shorter JORs.

However, Experiments 1–3 alone do not indicate whether participants can effectively monitor their learning with JORs with a within-subject manipulation that influences memory performance. As such, Experiment 4 was designed to determine whether participants’ JORs are sensitive to item difficulty, as previously reported for JOLs (e.g., Koriat et al., 2004; Kriot & Ma’ayan, 2005). In Experiment 4, item difficulty was manipulated by having participants study abstract and concrete words. Consistent with prior work (e.g., Begg, Duft, Lalonde, Melnick, & Sanvito, 1989; Hertzog, Dunlosky, Robinson, & Kidder, 2003; Paivio, 1966), it was expected that participants would recall more concrete words than abstract words. In addition, we anticipated that JORs would be shorter for abstract than for concrete words (cf. Begg et al., 1989; Hertzog et al., 2003).

Method

Participants

Thirty-four Colorado State University students ($M$ age = 18.94 years, $SD = 2.09$) received credit in psychology courses for participation.

Materials and procedure

Materials consisted of 30 words that varied on concreteness (range = 100–700; concrete words: $M = 631.46$, $SD = 9.43$; abstract words: $M = 258.23$, $SD = 10.90$) while controlling for word frequency (Kucera & Francis, 1967) and word length (word frequency: concrete words: $M = 37.62$, $SD = 28.47$; abstract words: $M = 44.08$, $SD = 50.18$; and word length: concrete words: $M = 5.85$, $SD = 0.48$; abstract words: $M = 5.85$, $SD = 1.21$). Two randomized versions of the study list
were created, consisting of 15 concrete words (e.g., rabbit) and 15 abstract words (e.g., value) that were randomly intermixed. Otherwise, the procedure was identical to that in Experiment 1.

Results

JORs and recall
Participants provided relatively short JORs (\(M = 15.28 \text{ min}, SD = 9.41 \text{ min}\)) that were reliably greater than zero, \(t(33) = 9.47\) (see Figure 1d). Moreover, JORs were sensitive to item type. Specifically, JORs for concrete words (\(M = 17.94, SD = 10.67\)) were significantly longer than JORs for abstract words (\(M = 12.74, SD = 9.06\)), \(t(33) = 4.91, \text{ Cohen’s } d = 0.53\). As expected, participants recalled a significantly greater percentage of concrete words (\(M = 39.28, SD = 18.82\)) than abstract words (\(M = 25.21, SD = 13.50\)), \(t(33) = 4.11, \text{ Cohen’s } d = 0.87\).

Resolution
A one-sample \(t\) test revealed that average gamma correlations (see Table 2) were significantly greater than zero, \(t(33) = 5.40\). Thus, when including item difficulty as a cue, JORs were sensitive to later recall. Further, correlations were reliably higher for concrete than for abstract items, \(t(33) = 2.06, \text{ Cohen’s } d = 0.49\). The correlation for concrete items reliably exceeded chance, \(t(31) = 3.71\), whereas the correlation for abstract items did not reliably differ from chance, \(t < 1\). These data indicate that JORs and recall were positively related and that participants were better than chance at determining which items would be retained for concrete but not abstract items.

Discussion
Experiment 4 replicated the prior experiments in demonstrating that participants were able to monitor their learning with JORs. Specifically, participants’ JORs were longer for concrete than for abstract items, consistent with the better memory evident for concrete items. Further, consistent with prior experiments, JORs were longer for items that were more likely to be remembered.

EXPERIMENT 5

Across the experiments reported thus far, average JORs were not extreme estimates of memory performance, as might be expected from the JOL literature (e.g., Koriat et al., 2004). Such data suggest that participants’ beliefs about retention may be more appropriately assessed via JORs than with JOLs, an issue we return to in the General Discussion. Further, Experiments 1–4 demonstrated that longer JORs were provided for items that were more likely to be remembered than for items that were later forgotten, indicating that JORs were related to memory performance. Experiment 4 also indicated that JORs were sensitive to item difficulty, because longer JORs were provided for concrete items than for abstract items. These data suggest that participants can monitor their learning with JORs in a similar manner to that with JOLs. Further, the underlying bases for JORs may be similar to those of JOLs. That is, JORs may be largely inferential and based on available cues during learning. While such data suggest some degree of convergence between JOLs and JORs, in other respects the judgements might rely on different information. We sought to examine such differences in Experiments 5 and 6 by investigating metacognitive control choices following each type of judgement.

In Experiments 5 and 6, we investigated whether JORs differentially influence study
decisions compared with other monitoring judgements. Specifically, participants in Experiment 5 were offered the opportunity to select items for restudy during encoding, although they did not actually restudy any items prior to test. Given that JORs required participants to consider future memory performance in units of time, such dimensions should be made salient to the participants. By extension, if time was made salient, then participants might consider the possibility of memory failure (e.g., a particular item might not be remembered after some duration) more so than in conditions that only required a probability judgement regarding the possibility that an item would be remembered. Thus, relative to JOLs, JORs may be more likely to induce participants to consider memory failure, thereby encouraging participants making JORs to select more items for restudy than would participants making JOLs. We tested this possibility in Experiment 5.

Participants in Experiment 5 were assigned to one of three conditions. One third of the participants made JOLs, one third made JORs, and the final third made no monitoring judgement (restudy only condition). Thus, the restudy only condition functioned as a baseline to compare study choices with the JOL and JOR conditions. Participants selected items to restudy in all conditions. For participants in the JOL and JOR conditions, the restudy decision immediately followed each judgement. Participants studied concrete and abstract items—as in Experiment 4—to examine the impact of item difficulty on restudy selection.

Method

Participants

One-hundred and twenty Colorado State University students (M age = 19.38 years, SD = 1.66) participated (40 in each condition) in exchange for course credit.

Design

A 2 (item type: abstract, concrete) × 3 (judgement condition: restudy only, JOL, JOR) mixed-factor design was employed with item type manipulated within subjects and judgement condition manipulated between subjects.

Materials and procedure

The materials used were identical to those in Experiment 4. Participants studied words presented one at a time for 4 s. One third of the participants made JORs indicating how long they would be able to remember each word on a minute scale (0–60 min). Another one third of the participants made JOLs predicting the likelihood of recalling each item on a later memory test on a scale from 0 (not likely at all) to 100 (very likely). Restudy choices immediately followed each JOR or JOL. In the restudy only condition, participants made the same restudy decision but with no intervening monitoring judgement. Participants in all conditions circled “YES” or “NO” on a sheet provided to indicate the choice to restudy a word. It should be noted that participants were not actually offered a restudy opportunity prior to free recall (cf. Finn, 2008) because our focal predictions pertained to the choices participants made and did not concern restudy per se. Following a 3-min filler task used in prior experiments, participants were given a 3-min free recall test.

Results

Recall

The mean percentage of items recalled was examined in a 2 (item type: abstract, concrete) × 3 (judgement condition: restudy only, JOL, JOR) mixed-factor ANOVA. Results showed that recall was significantly lower for abstract (M = 22.99, SD = 13.20) than for concrete words (M = 42.02, SD = 17.50), F(1, 117) = 140.97, ηp² = .55, replicating Experiment 4. There was no main effect for judgement condition, nor was there an Item Type × Judgement condition interaction (Fs < 1). Thus, participants in all conditions exhibited better memory for concrete than for abstract words.

Metamemory judgements

Due to the differing scales for JOLs (measured in percentages) and JORs (measured in minutes),
monitoring judgements were analysed separately. JOLs were submitted to a 2 (item type: abstract, concrete) × 2 (recall status: recalled, not recalled) repeated measures ANOVA (see Table 3). Analyses revealed that JOLs were higher for words that were recalled than for words that were not recalled, $F(1, 39) = 19.01, \eta^2_p = .33$. Further, JOLs were higher for concrete items than for abstract items, $F(1, 39) = 48.97, \eta^2_p = .56$. Finally, item type did not interact with recall status, $F(1, 39) = 1.71, p = .199, \eta^2_p = .04$.

JORs were also submitted to a 2 (item type: abstract, concrete) × 2 (recall status: recalled, not recalled) repeated measures ANOVA (see Table 3). Replicating Experiment 4, these analyses revealed that JORs were longer for words that were recalled than for words that were not recalled, $F(1, 39) = 14.60, \eta^2_p = .27$, and were longer for concrete items than for abstract items, $F(1, 39) = 35.51, \eta^2_p = .48$. Item type did not interact with recall status, $F(1, 39) = 2.89, p = .097, \eta^2_p = .07$. Taken together, these data indicate that participants’ judgements in both conditions discriminated between items based on item type and memory performance.

A comparison between the judgements in the judgement conditions (i.e., JOL condition, JOR condition) indicated that effect sizes did not reliably differ, $p = .80$. Thus, participants in both judgement conditions expected concrete words to be more memorable than abstract words.\(^3\)

**Restudy selection**

The mean percentage of items selected for restudy was analysed in 2 (item type: abstract, concrete) × 3 (judgement condition: restudy only, JOL, JOR) mixed-factor ANOVA (see Figure 2).\(^4\) Overall, significantly more abstract words ($M = 56.87, SD = 28.21$) were selected for restudy than concrete words ($M = 28.53, SD = 27.08$), $F(1, 117) = 188.27, \eta^2_p = .62$. A main effect of judgement condition was also present, $F(2, 117) = 3.71, \eta^2_p = .06$. In particular, participants in the JOR condition selected more items for restudy than those in the JOL condition, $t(78) = 2.03$, Cohen’s $d = 0.45$, and the restudy only condition, $t(78) = 2.21$, Cohen’s $d = 0.49$. However, there was no difference in the proportion of items selected for restudy between the JOL and restudy only conditions, $t < 1$. Finally, item type did not

\(^3\) To further compare the monitoring judgements, within-subjects $z$ scores were created for average JORs and JOLs. Results indicated that $z$ scores did not differ by condition, $F(1, 76) = 2.30, p = .134, \eta^2_p = .03$, and condition did not interact with item type, $F(1, 76) = 1.00, p = .320, \eta^2_p = .01$. These data suggest that participants’ judgements were of a similar magnitude for concrete and abstract items in the JOR and JOL conditions.

\(^4\) Several participants in Experiments 5 and 6 reported either invariant JOLs or invariant restudy choices. These participants were excluded from analyses, reflected in variations in the degrees of freedom reported for statistical tests in these experiments.
interact with judgement condition, $F < 1$. These data indicate that participants chose to restudy more items overall when making JORs than they did in the JOL and restudy only conditions.

Resolution
Gamma correlations were calculated based on the relation between judgement and recall, judgement and restudy, and restudy and recall and are described as follows. Gamma correlations for Experiments 5 and 6 are displayed in Table 4.

Judgement–recall gammas. A positive relationship was evident between judgements and recall for the JOL and JOR conditions, and these correlations were reliably greater than chance: JOL condition, $t(39) = 8.09$; JOR condition, $t(39) = 7.50$. However, correlations did not reliably differ between the JOL and JOR conditions, $t < 1$. Thus, resolution did not differ based on the type of judgement.

Judgement–restudy gammas. An inverse relationship was evident between judgements and restudy (i.e.,

Table 4. Kruskal–Goodman gamma correlations for Experiments 5 and 6

<table>
<thead>
<tr>
<th>Experiment</th>
<th>Condition</th>
<th>Correlation</th>
<th>Correlation</th>
<th>Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Judgement–recall</td>
<td>Judgement–restudy</td>
<td>Restudy–recall</td>
</tr>
<tr>
<td>Experiment 5</td>
<td>Restudy only</td>
<td>—</td>
<td>—</td>
<td>−.46 (.09)*</td>
</tr>
<tr>
<td>JOL</td>
<td>.37 (.05)*</td>
<td>−.91 (.02)*</td>
<td>−.41 (.09)*</td>
<td></td>
</tr>
<tr>
<td>JOR</td>
<td>.35 (.05)*</td>
<td>−.86 (.03)*</td>
<td>−.52 (.05)*</td>
<td></td>
</tr>
<tr>
<td>Experiment 6</td>
<td>JOF</td>
<td>.14 (.11)</td>
<td>−.37 (.26)</td>
<td>.20 (.29)</td>
</tr>
<tr>
<td>JOL</td>
<td>.35 (.09)*</td>
<td>−.83 (.05)*</td>
<td>−.05 (.29)</td>
<td></td>
</tr>
<tr>
<td>JOR</td>
<td>.35 (.09)*</td>
<td>−.80 (.11)*</td>
<td>.24 (.19)</td>
<td></td>
</tr>
</tbody>
</table>

Note: JOR = judgement of retention; JOL = judgement of learning; JOF = judgement of learning framed in terms of forgetting. The judgement type is specified by condition for each experiment. Standard errors provided in parentheses. Two correlations are not provided for the restudy only condition in Experiment 5 because monitoring judgements were not gathered in this condition. Judgements in the JOF condition in Experiment 6 were reverse scored for comparison purposes with the JOL conditions across experiments.

*Gamma significantly greater than chance (zero), $p < .05$. 
words with lower JOLs or shorter JORs were more likely to be selected for restudy), with means reliably different from zero for the JOL condition, \(t(32) = 46.78\), and JOR condition, \(t(38) = 31.62\). Follow-up tests indicated that there was no reliable difference between conditions, \(t(70) = 1.31, p = .19\). Thus, the relationship between judgements and restudy choices did not differ based on the type of judgement.

Restudy–recall gammas. An inverse relationship was found between restudy selection and recall for each judgement condition, and each correlation was reliably different from chance: JOL condition, \(t(32) = 4.58\); JOR condition, \(t(38) = 9.54\); and restudy only condition, \(t(36) = 5.06\). A one-way ANOVA indicated that there was no reliable difference between the restudy–recall correlations for the three conditions, \(F < 1\). These data indicate that participants selected to restudy words that were the least likely to be remembered.

Discussion

Participants in the JOR and JOL conditions in Experiment 5 successfully monitored their learning. That is, participants recalled significantly more concrete words than abstract words, and judgements (i.e., JORs and JOLs) were of a greater magnitude for concrete words than for abstract words (see also Experiment 4) and for words that were remembered than for words that were not remembered. Participants in all conditions chose to restudy significantly more abstract words than concrete words. Critically, participants who made JORs selected more items for restudy than did participants who made JOLs and participants who only made restudy choices. Thus, the monitoring judgement itself modified study choices. We suggest that JORs may bring to mind different information compared with JOLs or with no monitoring judgement. In particular, JORs may induce thoughts of memory failure, or may focus attention more carefully on the future memory test. We further explored the mechanism producing this effect in Experiment 6.

EXPERIMENT 6

Experiment 6 was designed to replicate and extend findings from Experiment 5. As previously noted, Finn (2008) reported that participants selected more items for restudy when judgements were framed in terms of forgetting than when they were framed in terms of remembering. Finn suggested that making judgements in terms of forgetting primed participants to consider memory failure, influencing their study decisions. If a similar mechanism drives monitoring with JORs, then participants’ restudy choices following a JOR may operate similarly to restudy choices following a JOL framed in terms of forgetting. That is, participants may select more items for restudy when making JORs than with a standard JOL condition, but restudy selection may not vary when making JORs compared with judgements of forgetting.

To explore this possibility, in Experiment 6 we used the same methodology as that of Experiment 5, with two notable exceptions. First, rather than using a restudy only condition, one group of participants made judgements of learning framed in terms of forgetting (JOFs). Thus, participants in Experiment 6 were assigned to one of three judgement conditions: JOR condition, JOL condition, or JOF condition. As previously noted, we anticipated that the proportion of items selected for restudy in the JOR condition might parallel the proportion of items selected for restudy in the JOF condition. Second, unlike in Experiment 5, participants were allowed to restudy items they selected for restudy in Experiment 6. Allowing participants to restudy items prior to test provided an indication of whether restudy decisions would impact later memory performance.

Method

Participants

Fifty-eight Kent State University students (\(M\) age = 20.24 years, \(SD = 2.33\); 19 in the JOF condition, 21 in the JOL condition, and 18 in the
JOR condition) participated in exchange for course credit.

**Design**

A 2 (item type: abstract, concrete) × 3 (judgement condition: JOF, JOL, JOR) mixed-factor design was employed with item type manipulated within subjects and judgement condition manipulated between subjects.

**Materials and procedure**

The materials used were identical to those in Experiments 4 and 5. In order to encourage participants to select items for restudy, participants studied words presented at a faster rate (2 s) than in Experiment 5. One group of participants made JORs (JOR condition) indicating how long they would be able to remember each word on a minute scale (0–60 min). Another group of participants made JOLs (JOL condition) predicting the likelihood of recalling each item on a later memory test on a scale from 0 (not likely at all) to 100 (very likely). Finally, a third group of participants made JOFs (JOF condition). Participants in the JOF condition made predictions indicating the likelihood of forgetting each item on a later memory test on a scale from 0 (not likely at all) to 100 (very likely). In all conditions, restudy choices immediately followed each judgement. Participants selected “YES” or “NO” on the computer screen to indicate the choice to restudy a word. Following a 3-min filler task used in Experiment 1, participants were re-presented with items selected for restudy in a new randomized order. Finally, participants were given unlimited time for free recall.

**Results**

**Recall**

The mean percentage of items recalled was examined in a 2 (item type: abstract, concrete) × 3 (judgement condition: JOF, JOL, JOR) mixed-factor ANOVA. Results showed that recall was significantly lower for abstract (M = 24.27, SD = 12.71) than for concrete words (M = 29.84, SD = 17.83), F(1, 55) = 6.09, ηp² = .02. There was no main effect for judgement condition, nor was there an Item Type × Judgement Condition interaction (Fs < 1). Thus, participants in all conditions exhibited better memory for concrete than for abstract words.

**Metamemory judgements**

Similar to Experiment 5, JOLs and JOFs (measured in percentages) were analysed together, but JORs (measured in minutes) were analysed separately. It should be noted that, following prior work (Finn, 2008), judgements in the JOF condition were reverse scored (e.g., a 40% JOF judgement was converted to a 60% JOL) in order to compare these judgements with JOLs. A 2 (item type: abstract, concrete) × 2 (judgement condition: JOF, JOL) mixed-factor ANOVA indicated that participants provided significantly higher JOLs for concrete words (M = 59.58, SD = 19.99) than for abstract words (M = 51.55, SD = 21.51), F(1, 38) = 20.46, ηp² = .35. There was no main effect for judgement condition, nor was there an Item Type × Judgement Condition interaction (Fs < 1). Thus, there were no differences in judgement magnitude between the JOL and JOF conditions. These data fail to replicate Finn (2008) who reported that JOFs were more conservative than JOLs (but see Kornell & Bjork, 2009; Rhodes & Castel, 2008).

Likewise, participants provided significantly longer JORs for concrete words (M = 25.80 min, SD = 17.57) than for abstract words, M = 19.28 min, SD = 12.70; t(17) = 2.87, Cohen’s d = 0.43. A comparison between the judgement conditions (i.e., JOF, JOL, JOR) indicated that effect sizes did not reliably differ, p = .52. Thus, participants in all judgement conditions expected concrete words to be more memorable than abstract words.

Metamemory judgements were then conditionized based on recall performance for each item type (see Table 3). JOLs and JOFs were submitted to a 2 (item type: abstract, concrete) × 2 (recall status: recalled, not recalled) × 2 (judgement condition: JOF, JOL) mixed-factor ANOVA (see Table 3). Analyses revealed that judgements were greater for words that were recalled than for...
words that were not recalled, $F(1, 37) = 8.71$, $\eta^2_p = .19$. Further, judgements were longer for concrete items than for abstract items, $F(1, 37) = 13.31$, $\eta^2_p = .27$. No other effects were supported, indicating that, in both conditions, judgement magnitude was greater for words that were recalled.

JORs were examined in a 2 (item type: abstract, concrete) $\times$ 2 (recall status: recalled, not recalled) repeated measures ANOVA, which revealed that JORs were marginally longer for words that were recalled than for words that were not recalled, $F(1, 17) = 4.38$, $p = .052$, $\eta^2_p = .21$. Further, JORs were longer for concrete items than for abstract items, $F(1, 17) = 6.22$, $\eta^2_p = .27$. Item type did not interact with recall status ($F < 1$), indicating that for both abstract and concrete items, longer JORs were provided for words that were recalled. Thus, participants’ judgements in all conditions discriminated between items based on item type and memory performance.5

Restudy selection

The mean percentage of items selected for restudy (Figure 3) was analysed in a 2 (item type: abstract, concrete) $\times$ 3 (judgement condition: JOF, JOL, JOR) mixed-factor ANOVA. Overall, significantly more abstract words ($M = 50.40$, $SD = 42.87$) were selected for restudy than concrete words ($M = 42.84$, $SD = 42.31$), $F(1, 55) = 16.90$, $\eta^2_p = .24$. A main effect of judgement condition was also present, $F(2, 55) = 3.92$, $\eta^2_p = .13$. In particular, participants in the JOL condition selected fewer items for restudy than did those in the JOF condition, $t(38) = 2.40$, Cohen’s $d = 0.76$, and the JOR condition, $t(37) = 2.50$, Cohen’s $d = 0.80$. However, there was no difference in the percentage of items selected for restudy between the JOF and JOR conditions, $t < 1$. Finally, item type did not interact with judgement condition, $F < 1$. These data indicate that participants chose to restudy more items overall when making JORs or JOFs than they did in the JOL condition.6

5To further compare the monitoring judgements, within-subjects $z$ scores were created for average JORs, JOLs, and reverse scored JOFs. Results indicated that $z$ scores did not differ by condition ($F$s < 1); further, condition did not interact with item type ($F < 1$) suggesting that participants’ judgements were of similar magnitude in all conditions.

6We note that providing a restudy opportunity did not influence memory performance for items selected for restudy compared with those items not selected for restudy, $F(1, 18) = 2.09$, $p = .165$, $\eta^2_p = .10$. 

Figure 3. Average percentage of items selected for restudy for the three judgement conditions in Experiment 6. JOF = judgement of learning framed in terms of forgetting; JOL = judgement of learning; JOR = judgement of retention. JOFs were reverse scored for comparison purposes with the JOL condition.
Resolution
Similar to Experiment 5, gamma correlations were calculated between judgements and recall, judgements and restudy, and restudy and recall (see Table 4). For the following analyses, judgements in the JOF condition were reverse scored (i.e., 100 – JOF).

Judgement–recall gammas. A positive relationship was evident between judgements and recall for the JOR, JOL, and JOF conditions. Further, correlations were reliably greater than chance for the JOR condition, $t(16) = 4.11$, and the JOL condition, $t(20) = 4.0$. However, correlations did not reliably differ from zero in the JOF condition, $t(18) = 1.39$, $p = 0.181$. A one-way ANOVA indicated that there was no reliable difference between judgement–recall correlations for the three conditions, $F(2, 54) = 1.60$, $p = 0.211$, $\eta^2_p = 0.06$. Thus, resolution did not differ based on the type of judgement.

Judgement–restudy gammas. An inverse relationship was evident between judgements and restudy (i.e., words with lower JOLs or shorter JORs were more likely to be selected for restudy) with means reliably different from zero for the JOL condition, $t(8) = 18.08$, and JOR condition, $t(8) = 7.94$. However, correlations were not reliably different than zero in the JOF condition, $t(9) = 1.50$, $p = 0.168$. A one-way ANOVA indicated that there was no reliable difference between judgement–restudy correlations for the three conditions, $F(2, 25) = 2.46$, $p = 0.106$, $\eta^2_p = 0.17$. Thus, the relationship between judgement and restudy choices did not differ based on the type of judgement.

Restudy–recall gammas. An inverse relationship was found between restudy selection and recall for each judgement condition; however, correlations were not reliably different from chance: JOL condition and JOF conditions, $t < 1$; JOR condition, $t(8) = 1.33$, $p = 0.219$. A one-way ANOVA indicated that there was no reliable difference between restudy–recall correlations for the three conditions, $F < 1$.

Discussion
Of primary interest, Experiment 6 demonstrated that study decisions differed following different metacognitive judgements even though monitoring accuracy was similar across judgement conditions. In particular, participants selected more items for restudy in the JOR condition than in the JOL condition (replicating Experiment 5). However, restudy selection did not differ between the JOR and JOF conditions, suggesting that the underlying mechanisms of these metacognitive judgements may be similar.

GENERAL DISCUSSION
With few exceptions, most work examining prospective memory monitoring has employed judgements of learning (JOLs). Thus, much of what is known about monitoring is primarily limited to one measure, JOLs. In an effort to expand understanding of monitoring, we reported a new measure, judgements of retention (JORs) that required participants to predict the amount of time information will be remembered. Evaluating monitoring with JORs is valuable to (a) evaluate whether the mechanisms that drive monitoring extend to time-based judgements, (b) assess participants’ predictions of memory retention, and (c) provide insight into those factors that influence metacognitive control processes. Several notable conclusions from the experiments provide insight into the underlying mechanisms of monitoring of learning with JORs.

Bases for monitoring with JORs
How is monitoring with JORs similar to monitoring with JOLs?
Participants can effectively monitor their learning with JORs (similar to monitoring measured with JOLs) because JORs were consistently related to memory performance. That is, in all experiments, longer JORs were observed for words that were actually remembered than for words that were not remembered. This effect obtained when providing participants with no scale for their JORs.
(Experiment 2), with a short RI (Experiments 1 and 2) or long RI (Experiment 3), with an item-level manipulation that influenced memory performance (Experiments 4, 5, and 6), and with restudy decisions (Experiments 5 and 6). Further, within-subjects gamma correlations indicated that JORs and recall were positively related across all experiments, and participants’ JORs were also sensitive to manipulations that influenced memory performance (Experiments 4, 5, and 6). Finally, Experiments 5 and 6 demonstrated that the item level relationships (i.e., resolution) between JORs, recall, and restudy mirrored those of the JOL condition, replicating prior work (e.g., Dunlosky & Thiede, 2004; Nelson, Dunlosky, Graf, & Narens, 1994; Thiede & Dunlosky, 1999). Thus, similar to monitoring of learning with JOLs, JORs can be based on factors that influence memory performance.

JORs are likely to be inferential in nature, rather than driven by direct access to memory strength. An inferential account would suggest that JORs are based on salient information during learning, which may, or may not, be diagnostic of future memory performance (see Koriat, 1997, for a useful framework). For example, the current research indicates that participants attend to cues such as item difficulty with JORs in similar ways to those that have been demonstrated with JOLs (e.g., Koriat et al., 2004; Koriat & Ma’ayan, 2005). However, some inconsistencies were apparent between predictions and actual memory performance (e.g., some items were given short predictions, but were remembered, and other items were given long predictions, but were forgotten). Thus, we suggest that JORs are a comparative, cue-based measure of monitoring similar to JOLs.

**How is monitoring with JORs dissimilar from monitoring with JOLs?**

Each experiment demonstrated that JORs were not extreme estimates of memory performance. Such memory durations contrast with what may be expected from when judgements are made using JOLs (e.g., Carroll et al., 1997; Koriat et al., 2004; Shaddock & Carroll, 1997). That is, as previously noted, prior research has shown that monitoring can be poor with JOLs. In particular, Koriat et al. (2004) evaluated participants’ JOLs and memory performance for immediate recall, recall in 1 day, or recall in 1 week. Results indicated that JOLs were approximately the same for all three retention intervals, even though memory performance declined with longer retention intervals. Koriat et al. (2004, Experiment 3B) evaluated participants’ JOLs further using a within-subjects manipulation of RI such that participants made predictions for each of three RIs: 10 min, 1 day, and 1 week. Results indicated that JOLs were sensitive to RI because JOLs were lower for the longer RIs. However, average JOLs were 67.18% for a 10-min delay, 57.09% for a 1-day delay, and 48.84% for a 1-week delay, and memory performance (in an earlier experiment) averaged 52.60% for a 10-min delay, 28.88% for a 1-day delay, and 17.96% for a 1-week delay. Thus, even when a within-subjects manipulation of RI was used, participants’ JOLs did not correspond with memory performance and were overconfident.

In contrast with JOLs, JORs provide a different assessment of participants’ beliefs about retention. In the current research, across six experiments, average JORs were relatively short estimates of memory duration (compared with prior research using JOLs). That is, on average, participants indicated that they would remember information for approximately 15 min (Experiments 1 and 4), 7 min (Experiment 2), 18 min (Experiment 3), or 22 min (Experiments 5 and 6). Thus, compared with JOLs, JORs reflect more realistic ideas about memory retention. Indeed, participants in the current experiments identified very few items that would be remembered at long delays. For example, in Experiment 2 (the experiment with the most extreme JORs), only 3% of items were given JORs of one day or greater, and in Experiment 6 (the experiment with the longest average JORs), only 13.5% of items were given JORs of 1 hour. Thus, in contrast to the JOL literature, the retention estimates provided with JORs indicate that participants do not expect to remember such high percentages of items after long delays.

Other contrasts with JOLs are also evident. For example, when JORs are made, each item is assigned a time indicating the durability of
memory and at what point forgetting is likely to occur for that particular item. We examined this in Experiment 6 by comparing JORs with JOLs framed in terms of forgetting (JOFs). The results showed that restudy decisions did not differ between the JOR and JOF conditions. Thus, although JOFs may not always provide differences in judgement magnitude relative to JOLs, JORs and JOFs may bring memory failure to mind, evident in the larger number of items selected for restudy than in the standard JOL condition. However, JORs provide a measure of memory duration that is not available with standard either JOLs or JOFs. In addition, JORs led to different patterns of restudy selection compared with standard JOLs. In particular, participants who made JORs chose more items for restudy than did participants in the JOL condition (Experiments 5 and 6) or a restudy only condition (Experiment 5). Such data indicate that JORs may be equally suitable to measure metacognitive monitoring as are JOLs, while providing a context in which different choices are made, perhaps by encouraging participants to consider different information when making memory predictions.

As such, the JOR scale might prime theories of forgetting, bringing to mind instances when memory failed over longer time intervals (Finn, 2008; Koriat et al., 2004). However, compared with JOLs, by making time salient via framing metacognitive judgements as judgements of retention (i.e., JORs), JORs may be based on cues available that may be predictive of future memory performance rather than on the current status of encoding, providing a context in which participants more carefully consider the impending test. Alternatively, the discrepancy-reduction model (e.g., Thiede & Dunlosky, 1999) posits an interaction between control and monitoring (i.e., JOL or JOR) processes such that efforts will be made to continue to learn material until the performance goal is likely to be reached (but see Ariel, Dunlosky, & Bailey, 2009; Metcalfe & Kornell, 2005, for alternatives). Thus, the JOR data could be interpreted as enhancing awareness (at least implicitly) of the general discrepancy between learning and memory performance, leading participants to more frequently seek to restudy items. It is also possible that making JORs, as opposed to JOLs, produced a different goal for study. That is, JORs may alter the goal for study relative to JOLs, perhaps placing a stronger emphasis on mastery, thus leading to differences in the study choices that are made. This suggestion would support prior work demonstrating that manipulations of a study goal (i.e., an easy goal versus a more challenging goal) modify study choices (e.g., Thiede & Dunlosky, 1999). These alternative hypotheses for the differential study decisions following JORs warrant further investigation.

Future directions

The current experiments also suggest directions for future research to further evaluate memory monitoring with JORs. For example, whereas memory is probabilistic (i.e., information can be forgotten after a 10-min delay, but then remembered 45 min later) JORs are made in discrete units of time. One potential method of introducing a probabilistic component to JORs would be to assess second-order judgements (SOJs; Dunlosky, Serra, Matvey, & Rawson, 2005; see also Miller & Geraci, 2011). For example, Dunlosky et al. (2005) had participants make JOLs and SOJs (concerning the likelihood that a JOL was correct) for pairs of words. Their data showed dissociations between JOLs and SOJs, with the highest and lowest JOLs generally accompanied by the highest levels of confidence that the judgement was correct. By extension, measuring monitoring via JORs and collecting SOJs so as to capture confidence in the correctness of the specified memory duration might provide further insight into participants’ beliefs about memory retention. For example, lower SOJs might indicate that JORs were made with less confidence, and these items might have a lower probability recall on a future test.

Further, the reported experiments demonstrate that the scale provided for JORs influenced memory durations (i.e., Experiments 1 and 2). However, it is unclear what influence such scaling issues may have on restudy decisions. For example, a short-duration scale (e.g., JOR scale ranging from 0 to 5 min) may create a false sense of mastery such that the number of items selected...
for restudy would not vary between a JOR condition and a JOL condition. Thus, future research should focus on further evaluating the impact of JORs on study decisions.

CONCLUSIONS

The data reported evaluated the underlying mechanisms of monitoring of learning with an entirely new measure of monitoring, judgements of retention (JORs). This work offers a unique measure of monitoring that provides an item-by-item index of assessments of memory durability, information not captured by JOLs. Thus, the JOR technique provides further insight into participants’ awareness of forgetting that occurs over time and provides extra methodological power for researchers interested in investigating people’s beliefs about retention. Further, JORs provide a context in which different study choices are made, suggesting that considering the durability of a memory may provide a different perspective on memory monitoring compared with other measures.

REFERENCES


