Does the amount of material to be remembered influence judgements of learning (JOLs)?

Sarah K. Tauber and Matthew G. Rhodes
Colorado State University, Fort Collins, CO, USA

The current study examined predictions of memory performance as a function of the amount of information to be remembered. In four experiments participants studied and made judgements of learning (JOLs) for long or short lists of words. Results demonstrated that participants provided lower JOLs for long compared with short lists. However, whereas JOLs for short lists strongly corresponded with memory performance, participants’ JOLs were consistently overconfident for long lists. Participants were unable to remedy this overconfidence for long lists even when provided information about the list length conditions or warned that a long list of words is difficult to learn. Only when given a prior list learning experience were JOLs for a long list consistent with memory performance. These data indicate that predictions of memory performance are sensitive to the amount of material TBR. However, predictions only correspond with the amount of information to-be-remembered under limited circumstances, providing support for frameworks which suggest that memory predictions are inferential in nature.

Keywords: Metacognition; Metamemory; Memory; List length.

Do people account for the amount of material to be remembered (TBR) when predicting future memory performance? For example, consider students preparing for an upcoming examination. Will they alter their study habits if the exam covers material from three chapters rather one chapter? While memory performance should decline as the amount of material TBR increases (cf. Ebbinghaus, 1913; Strong, 1912), researchers have yet to explore whether people are aware of this effect. Thus in the current study we examined the degree to which participants are aware, and the means by which they can be made aware, of the impact of the amount of material TBR on subsequent memory performance. In particular, we focus on two specific questions. Are participants sensitive to the amount of material TBR?

In addition, does the overall correspondence (calibration) between predictions and memory performance vary based on the amount of material TBR?

Prior studies examining predictions of memory performance have typically solicited judgements of learning (JOLs) either immediately after the presentation of an item or following a delay (for reviews see Koriat, 2007; Metcalfe, 2000). A number of studies (e.g., Koriat, 1997; Koriat, Bjork, Sheffer, & Bar, 2004; Moulin, Perfect, & Jones, 2000) have explored the degree to which JOLs differ based on the conditions present during study (i.e., metacognitive sensitivity). For example, Moulin et al. compared metacognitive sensitivity for healthy older adults and an Alzheimer’s disease (AD) impaired sample by soliciting

© 2010 Psychology Press, an imprint of the Taylor & Francis Group, an Informa business
http://www.psypress.com/memory
DOI:10.1080/09658211003662755
JOLs for a set of easy and difficult words. Results demonstrated that while healthy older adults exhibited better memory performance than individuals with AD, both groups were sensitive to item difficulty, providing lower JOLs for difficult compared with easier items.

In contrast, absolute accuracy refers to the extent to which participants’ predictions are calibrated (i.e., correspond) with subsequent performance. Thus, whereas metacognitive sensitivity focuses solely on predictions, calibration reflects how well predictions match actual memory performance. While JOLs are often accurate, a number of important departures have been observed between actual and predicted memory performance (e.g., Begg, Duft, Lalonde, Melnick, & Sanvito, 1989; Benjamin, Bjork, & Schwartz, 1998; Castel, McCabe, & Roediger, 2007; Koriat & Bjork, 2005; Mazzoni & Nelson, 1995; Rhodes & Castel, 2008a, 2009). For example, Koriat et al. (2004) had participants study related and unrelated word pairs and make immediate JOLs regarding the probability of recalling the second word of the pair either immediately, after 1 day, or after 1 week. Results showed that whereas recall was far better when tested immediately rather than 1 day or 1 week later, JOLs remained constant across retention intervals. Such discrepancies provide some indication of the bases for JOLs and information that is or is not attended to when making predictions.

 Likewise, and of relevance to the current study, several prior studies have reported discrepancies between memory predictions and recall under conditions of interference (e.g., Metcalfe, Schwartz, & Joaquin, 1993; Rhodes & Castel, 2008b). For example, after studying a list of words, Rhodes and Castel (2008b) had participants either engage in a free-recall task or attempt to recall the study list in the context of a randomly selected subset of the study list (i.e., part-set cueing). While recall was proportionately lower when participants were exposed to part-set cues (cf. Roediger, Stellon, & Tulving, 1977; Slamecka, 1968) predictions were similar for the part-set cueing and free recall conditions. Thus, participants’ predictions did not accurately account for the interference wrought by part-set cues. However, other researchers (e.g., Maki, 1999) have observed that predictions can account for variables that engender interference. For example, Maki had participants make memory predictions for number–word pairs (e.g., 261–farmer) during an initial study session. Participants were then presented with pairs consisting of the same number with new words (e.g., 261–hunter) and entirely new number–word pairs (e.g., 345–garden). Finally, memory for the number–word pairs from the first session was tested. Maki (1999) reported that JOLs and recall were lower for cues (i.e., 261–) presented with two words across encoding sessions (e.g., 261–farmer, 261–hunter) than for entirely new pairs (e.g., 345–garden). Thus memory predictions were sensitive to interference and were well calibrated with the interference produced by studying different words with the same cue. In the current study, long lists of words may produce high levels of interference compared with shorter lists of words. The mixed evidence in the literature regarding metacognitive awareness of interference begs the question of whether participants’ predictions will accommodate the impact of interference in this context.

A direct access hypothesis (e.g., Arbuckle & Cuddy, 1969; Hart, 1965; Dunlosky & Nelson, 1994) can account for the possibility that memory predictions are sensitive to interference (cf. Maki, 1999). According to this hypothesis, participants assess the strength of memory traces for each event and JOLs are based on these assessments of strength. Thus metacognitive sensitivity would be evident if low JOLs were provided for a weak memory trace and high JOLs were provided for a strong memory trace. Likewise, according to the direct access view, JOLs would be well calibrated when weaker memory traces elicit lower JOLs (and lower subsequent memory performance) than stronger memory traces. Based on this account, memory strength should be weaker for long lists and thus beget low JOLs and poorer memory performance. Conversely, shorter lists should yield a stronger memory trace and thus higher JOLs that accompany better memory performance.

In contrast, Koriat (1997) has suggested that memory predictions reflect an inferential process that relies on the cues available to the rememberer rather than direct assessments of memory strength. Koriat’s (1997) cue-utilisation framework holds that three types of cues influence predictions: intrinsic, extrinsic, and mnemonic cues. Intrinsic cues involve characteristics of TBR items that may be diagnostic of ease of learning (e.g., relatedness, imageability). Extrinsic cues include the learning conditions present and modified by the participant (e.g., retention interval, type of encoding strategy employed). Finally, mnemonic cues are internal
indices that inform learning of a particular item (e.g., practice, experience). Koriat’s framework thus suggests that participants will demonstrate metacognitive awareness to the extent that the cues that are attended to are likewise diagnostic of performance. As noted previously, prior work (e.g., Koriat et al., 2004) suggests that people may attend to intrinsic cues (e.g., qualities of the items studied) at the expense of other information (e.g., retention interval) that is diagnostic of performance. Thus, the cue-utilisation framework readily accounts for the existing data demonstrating poor calibration for interference effects (e.g., Rhodes & Castel, 2008b). For the current study, if participants’ predictions are indeed unable to account for the impact of list length, this would suggest that predictions rely on the potential cues available rather than direct access to the strength of a memory trace.

THE CURRENT STUDY

The current study explored metacognitive sensitivity and (absolute) metacognitive accuracy as a function of the amount of material TBR. In four experiments participants made item-by-item JOLs for words in a long list (100 words; 60 words) or short list (10 words). Experiment 1 examined the impact of list length on memory predictions and recall. To anticipate, results from this experiment showed that while participants were sensitive to list length (i.e., they provided lower JOLs for the short list compared with the longer list), JOLs were poorly calibrated for long lists and well calibrated for short lists. The remaining experiments attempted to focus participants’ attention on list length in an effort to improve calibration for long lists. Thus, participants in Experiment 2 were provided with information about all list length conditions (cf., Koriat et al., 2004). Participants in Experiment 3 were given an explicit warning indicating that learning longer lists negatively impacts memory performance. Finally, Experiment 4 employed a within-participants manipulation wherein participants studied a long list followed by a short list of words or vice versa. To our knowledge, the current study is the first to examine participants’ awareness of the impact of the amount of information TBR and those conditions that may foster awareness of the effects of list length on memory performance.

EXPERIMENT 1

Half of the participants in Experiment 1 were asked to make JOLs for either a long list (100 words) or short list (10 words) of words followed by a recall test. If participants are sensitive to list length, then JOLs should be lower for the long list compared with the short list. If participants are not sensitive to list length, then JOLs should not differ between the list length conditions. Based on prior work (e.g., Ebbinghaus, 1913; Strong, 1912) we anticipated that participants would remember proportionately fewer words from the long list than the short list. Thus participants will be well calibrated to the extent that their memory predictions correspond to that level of memory performance. However, a different pattern of calibration between conditions would suggest that participants are unable to adequately account for the impact of list length on memory performance.

Method

Participants. A total of 32 individuals from Colorado State University participated (M age = 20.72, SD = 3.04) and received course credit. All participants were tested individually in this and subsequent experiments.

Materials. A 100-item word list was created consisting of nouns (e.g., palace, circle, moon) equated on Kucera-Francis word frequency (M = 37.24, SD = 13.60; MRC Psycholinguistic Database, 1987), number of letters (M = 4.91, SD = 1.12), and number of syllables (M = 1.48, SD = 0.62). From this 100-item word list, 10-item word lists were created such that each set of words served equally often in the 100-word or 10-word condition.

Procedure. After providing consent participants were informed of how many words they would be asked to remember (i.e., either 100 or 10 words). Words were then presented one at a time at a 4-second rate, from either the long or short list, and participants were instructed to learn the words in preparation for a later memory test. Immediately following the presentation of each word participants were given 5 seconds to rate the likelihood of later recalling that word on a scale from 0 (not likely at all) to 100 (very likely). Participants wrote each JOL on an answer sheet provided and were encouraged to use the entire range of the
scale. A 500-ms interstimulus-interval (ISI) preceded the presentation of the next word for study. Following a 3-minute filler task (writing down states of the United States) participants were allotted 3 minutes for free-recall.

**Results**

**JOL sensitivity.** Predicted and actual recall performance are displayed in Figure 1. The sensitivity of JOLs to list length was first analysed with a one-way ANOVA comparing JOLs for the 10-item and 100-item conditions. These data showed that JOLs were reliably higher in the 10-item condition compared with the 100-item condition, $F(1, 30) = 21.17$, Cohen's $d = 1.68$.

**Calibration.** Predicted and actual recall performance were analysed in a 2 (List Length: 10-item, 100-item) × 2 (Measure: JOL, recall) mixed-factor analysis of variance (ANOVA). Main effects of Measure, $F(1, 30) = 25.21$, $\eta^2_p = .46$, and List Length, $F(1, 30) = 53.88$, $\eta^2_p = .64$, were qualified by a significant Measure × List Length interaction, $F(1, 30) = 4.54$, $\eta^2_p = .13$. Follow-up tests indicated that participants' JOLs in the short-list length condition were well calibrated with no reliable difference between JOLs and recall, $t(15) = 1.77$, $p = .10$, Cohen's $d = 0.60$. In contrast, participants' JOLs were reliably greater than recall performance in the long-list length condition, $t(15) = 6.23$, Cohen's $d = 2.23$.

**Item position effects.** While participants given a long list exhibited reliably poorer calibration than participants given a short list, calibration may change across the longer list. For example, as proactive interference builds up, participants may provide lower JOLs for items at later list positions. Thus, calibration may be better for items presented later rather than earlier in a long list. In order to assess whether calibration changed across items we calculated average JOLs and recall for the first 10 words and for the last 10 words in the 100-item condition (see the top row of Table 1). These data were examined in a 2 (Measure: JOL, recall) × 2 (List Position: first 10, last 10) repeated-measures ANOVA. Overall, JOLs exceeded recall performance, $F(1, 15) = 38.05$, $\eta^2_p = .72$. However, there was no main effect of List Position nor did List Position interact with Measure, $F$s < 1. Thus there was no significant change in calibration from the first 10 items to the last 10 items. In addition, a comparison of JOLs for the first 10 items and last 10 items indicated that JOLs did not reliably vary from the beginning to the end of the list, $t(15) = 1.76$, $p = .10$. Therefore, JOLs were neither better calibrated nor more sensitive to list length by the end of the 100-item word list.

**EXPERIMENT 2**

Results from Experiment 1 showed that participants given a short list provided lower JOLs than participants given a long list, suggesting that JOLs are sensitive to list length. However, while JOLs strongly corresponded with recall performance in the short-list condition, calibration was poor in the long-list condition, as JOLs far exceeded their recall performance. Thus, calibration appears to suffer when large amounts of material must be learned. Data from Experiment 1 beg the question of whether calibration can be improved for long lists. In particular, JOLs in Experiment 1 may have been driven largely by aspects of the word itself (an intrinsic cue in Koriat’s 1997 cue-utilisation framework) at the expense of cues related to the features of the study list, such as its length (an extrinsic cue). Therefore, ameliorating calibration may require a manipulation that draws participants’ attention to the study conditions rather than to the TBR item itself.

To this end, Experiment 2 attempted to make list length a more salient cue. In particular, participants were informed about both list lengths presented later rather than earlier in a long list. In order to assess whether calibration changed across items we calculated average JOLs and recall for the first 10 words and for the last 10 words in the 100-item condition (see the top row of Table 1). These data were examined in a 2 (Measure: JOL, recall) × 2 (List Position: first 10, last 10) repeated-measures ANOVA. Overall, JOLs exceeded recall performance, $F(1, 15) = 38.05$, $\eta^2_p = .72$. However, there was no main effect of List Position nor did List Position interact with Measure, $F$s < 1. Thus there was no significant change in calibration from the first 10 items to the last 10 items. In addition, a comparison of JOLs for the first 10 items and last 10 items indicated that JOLs did not reliably vary from the beginning to the end of the list, $t(15) = 1.76$, $p = .10$. Therefore, JOLs were neither better calibrated nor more sensitive to list length by the end of the 100-item word list.

![Figure 1. Experiment 1 calibration data.](image-url)
used in the experiment prior to studying their own particular list (either a long or short list). Critically, this information was reiterated and emphasised by directing participants to be mindful of these conditions while making predictions. Participants may be well calibrated for long list length conditions, but only when provided with such relative information about the amount of material TBR (cf. Koriat et al., 2004). For example, Koriat (1997) has suggested that memory predictions reflect one’s personal theories about memory (i.e., theory-based knowledge) or one’s familiarity with a particular task garnered through experience (i.e., experience-based knowledge). Accordingly, providing participants with information about the list length conditions should allow them to make more informed theory-based memory predictions in comparison with Experiment 1. If describing and emphasising both list length conditions indeed improves calibration, then JOLs for participants in the 100-word condition should be as well calibrated with recall performance as JOLs for participants in the 10-word condition.

**Method**

**Participants.** A total of 32 individuals from Colorado State University participated (M age = 19.28, SD = 1.69) and received course credit.

**Materials and procedure.** The materials and procedure were identical to those used in Experiment 1 with one exception. Specifically, participants were informed that some participants in the experiment would be learning 100 words while others would be learning 10 words. Participants were additionally instructed to keep these varying list lengths in mind while making predictions for later memory performance.

**Results**

**JOL sensitivity.** Mean JOLs and the mean percentage of items recalled in Experiment 2 are presented in Figure 2. A one-way ANOVA comparing JOLs between the two list length conditions (i.e., 10-item, 100-item) indicated that JOLs were reliably higher in the 10-item condition compared with the 100-item condition, $F(1, 30) = 4.72$, Cohen’s $d = 0.79$, consistent with Experiment 1.

**Calibration.** Predicted and actual recall performance was analysed in a 2 (Measure: JOLs, recall) x 2 (List Length: 10 words, 100 words) mixed-factor ANOVA. Reliable main effects of Measure, $F(1, 30) = 20.66$, $\eta_p^2 = .41$, and List Length, $F(1, 30) = 19.30$, $\eta_p^2 = .39$, were qualified by a significant Measure x List Length interaction, $F(1, 30) = 4.58$, $\eta_p^2 = .13$. Replicating Experiment 1, follow-up tests showed that JOLs were reliably higher than recall in the 100-item list condition, $t(15) = 4.89$, Cohen’s $d = 1.55$, while no reliable difference between JOLs and recall was evident for the 10-item list length condition, $t(15) = 1.65$, $p = .12$.

These data suggest that the condition comparison did not improve calibration for individuals in the long-list condition relative to participants who received a short list. However, the condition information may have had some impact on calibration for participants in the long-list condition if compared with a group of participants who did not receive any information. We did not

---

**TABLE 1**

<table>
<thead>
<tr>
<th></th>
<th>Item position data for Experiments 1-4</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>First 10 items</td>
</tr>
<tr>
<td></td>
<td>JOL</td>
</tr>
<tr>
<td>Experiment 1</td>
<td>47.34 (17.17)</td>
</tr>
<tr>
<td>Experiment 2*</td>
<td>54.48 (25.47)</td>
</tr>
<tr>
<td>Experiment 3*</td>
<td>41.57 (19.87)</td>
</tr>
<tr>
<td>Experiment 4, Time 1</td>
<td>40.22 (15.38)</td>
</tr>
<tr>
<td>Experiment 4, Time 2</td>
<td>38.59 (15.88)</td>
</tr>
</tbody>
</table>

Standard deviations provided in parentheses. *reliable difference in JOLs.
include such a condition in Experiment 2, but note that we used the same procedure (without condition information) and identical materials in Experiment 1.

Thus we further examined the potential impact of additional condition information in a 2 (Measure: JOL, recall) × 2 (Experiment: 1, 2) mixed-factor ANOVA examining data for participants given a 100-word list. For this analysis, Experiment 1 served as a no-information control condition and Experiment 2 served as a condition-comparison condition. While JOLs exceeded recall, $F(1, 30) = 59.30, \eta^2_p = .66$, there was no main effect of Experiment, nor an Experiment × Measure interaction, $Fs < 1$. This subsidiary analysis suggests that the condition information led to no difference in JOLs or recall for Experiment 2 compared with the control condition in Experiment 1.

**Item position effects.** As in Experiment 1, we examined calibration as a function of list position with a 2 (Measure: JOL, recall) × 2 (List Position: first 10, last 10) repeated-measures ANOVA for the 100-item condition (see Table 1). A main effect of Measure was evident, $F(1, 15) = 24.37, \eta^2_p = .62$, as was a main effect of List Position, $F(1, 15) = 6.27, \eta^2_p = .30$. The interaction between Measure and List Position was not reliable, $F(1, 15) = 1.49, p = .24, \eta^2_p = .09$. Replicating Experiment 1, there was no significant change in calibration from the first 10 items to the last 10 items. However, in contrast to Experiment 1, participants’ JOLs reliably decreased from the first 10 items to the last 10 items, $t(15) = 4.28, Cohen's d = 2.21$. Thus the condition comparison manipulation reduced JOLs by the end of the 100-item list, although it was not similarly effective in improving calibration for items at the end of the list.

### EXPERIMENT 3

Experiment 2 replicated Experiment 1 such that JOLs were lower for the long-list compared with the short-list condition. However, Experiment 2 demonstrated that informing participants about the different list lengths used was not sufficient to improve calibration for the 100-word condition (cf. Koriat et al., 2004). That is, even when participants were informed that they would receive either a 10-word or 100-word list, and when likewise instructed to keep this information in mind when making JOLs, participants continued to exhibit the poorest calibration when studying 100 words. Experiment 3 thus attempted to make list length information more salient by specifically warning participants that a long list of words can be difficult to remember. Such a warning may better emphasise theory-based knowledge of list length in comparison to the first two experiments, and may improve calibration for the 100-word condition. However, we note that an improvement in calibration is not inevitable. For example, Rhodes and Castel (2008a) had participants make JOLs for words presented in a large or small font size. While font size had no impact on memory performance, JOLs were significantly higher JOLs for large compared with small words. This metacognitive illusion persisted, although diminished in magnitude, even when participants were warned that font size would have no bearing on memory performance. In a similar sense, participants may continue to exhibit poor calibration for a long list even when specifically warned about its detrimental effects on memory performance.

### Method

**Participants.** A total of 32 individuals from Colorado State University participated ($M$ age = 19.25, $SD = 0.98$) and received course credit.

**Materials and procedure.** The materials and procedure were identical to those used in Experiment 1, with the exception that participants were provided with a warning about the nature of the list length effect. In particular, participants were told:

In this experiment some people are asked to learn 100 words and other people are asked to learn 10 words. It has been shown (in previous
that learning a long list of words (100 words) is much harder than learning a short list of words (10 words), so please keep this in mind when making your predictions.

**Results**

**JOL sensitivity.** Mean JOLs and the mean percentage of items recalled in Experiment 3 are presented in Figure 3. A one-way ANOVA indicated that JOLs were reliably higher in the 10-item condition compared with the 100-item condition, $F(1, 30) = 24.21$, *Cohen’s $d = 1.80$.

**Calibration.** Predicted and actual recall performance were analysed in a 2 (List Length: 10 words, 100 words) x 2 (Measure: JOL, recall) mixed-factor ANOVA. A main effect of Measure, $F(1, 30) = 1.76$, *$p = .20$, $\eta_p^2 = .06$, was not supported, but a reliable main effect of Condition, $F(1, 30) = 66.35$, *$\eta_p^2 = .69$, was qualified by a significant Measure x Condition interaction, $F(1, 30) = 5.66$, *$\eta_p^2 = .16$. Replicating the previous experiments, follow-up tests demonstrated that JOLs were significantly higher than recall for the long-list condition, $t(15) = 3.91$, *Cohen’s $d = 1.26$, while no reliable difference was evident between JOLs and recall for the short-list length condition ($t < 1$).

These data suggest that the warning did not improve calibration for individuals in the long-list condition relative to participants who received a short list. We further examined the potential impact of a warning in a 2 (Measure: JOL, recall) x 2 (Experiment: Experiment 1, Experiment 3) mixed-factor ANOVA examining data for participants given a 100-word list. For this analysis, Experiment 1 served as a no-warning control condition and Experiment 3 served as a warning condition. Results revealed reliable main effects of Measure, $F(1, 30) = 53.31$, *$\eta_p^2 = .64$, and Experiment, $F(1, 30) = 5.59$, *$\eta_p^2 = .16$. These main effects were qualified by a reliable Measure x Experiment interaction, $F(1, 30) = 5.66$, *$\eta_p^2 = .16$. Specifically, whereas participants in both experiments had equivalent levels of recall ($t < 1$), participants who were warned (i.e., those in Experiment 3) exhibited reliably lower JOLs than participants who had not been warned (i.e., those in Experiment 1), $t(30) = 2.65$, *Cohen’s $d = 0.97$. Thus this subsidiary analysis suggests that the warning led to more conservative predictions but was not sufficient to fully remedy calibration when participants studied a long list.

**Item position effects.** As in prior experiments, we examined performance in the 100-item condition as a function of item position in a 2 (Measure: JOL, recall) x 2 (List Position: first 10, last 10) repeated-measures ANOVA (see Table 1). Reliable main effects of Measure, $F(1, 15) = 14.81$, *$\eta_p^2 = .50$, and List Position, $F(1, 15) = 6.84$, *$\eta_p^2 = .31$, were qualified by a marginally reliably Measure x List Position interaction with List Position, $F(1, 15) = 3.96$, *$p = .07$, *$\eta_p^2 = .21$. Follow-up tests demonstrated a significant reduction in JOLs from the first 10 to the last 10 items, $t(15) = 2.59$, *Cohen’s $d = 1.34$, while no difference in recall was evident, $t < 1$. Thus, our warning manipulation marginally improved calibration by the end of the 100-item list due to a reduction in JOLs.

**EXPERIMENT 4**

Experiment 3 replicated previous experiments demonstrating metacognitive sensitivity to list length, as JOLs were lower for the long list compared with the short list. However, participants were more poorly calibrated for long compared with short lists even when warned in Experiment 3. In conjunction with data from Experiment 2 showing that a comparison of conditions failed to improve calibration, these data suggest that participants have difficulty incorporating information about list length (an extrinsic cue) into their JOLs. It may be that participants lack theory-based knowledge about list length, or that such theories may be particularly challenging to evoke. Prior work indicates that for some extrinsic cues, a within-participants manipulation may be necessary in order to

![Figure 3](image-url). Experiment 3 calibration data.
improve metacognition by allowing participants to rely on experience-based knowledge (e.g., Koriat et al., 2004). For example, Koriat et al. (2004; see also Rhodes & Castel, 2008b) observed that a within-participants manipulation was sufficient for participants to attend to retention interval information, such that participants were better able to account for retention interval when asked to make predictions for each of three retention intervals (i.e., 10 minutes, 1 day, and 1 week).

To this end, Experiment 4 employed a within-participants manipulation that gave participants experience in learning a long and short list of words. Specifically, participants either first learned a short list of words (10 words) followed by a longer list of novel words (60 words) or first learned a long list followed by a short list of words.\footnote{A list length of 60 words was selected for the longer list-length condition in order to be able to fit the experiment into a 1-hour session and to explore differing list length conditions across experiments.} Based on prior work (e.g., Koriat et al., 2004; Rhodes & Castel, 2008b), we anticipated that a within-participants manipulation of list length would increase participants’ awareness of the detrimental impact of a long list on memory at Time 2, such that participants would be better calibrated for long lists. In other words, calibration at Time 1 should replicate prior experiments, with participants better calibrated for the short-list condition compared with the long-list condition. However, at Time 2, calibration was expected to improve for the long-list condition.

**Method**

**Participants.** A total of 28 individuals from Colorado State University participated ($M_{\text{age}} = 19.07$, $SD = 1.96$) and received course credit.

**Materials and procedure.** Identical materials from Experiment 1 were used, with slight changes in the word list and procedure. Two 60-item word lists were created, one for the Time 1 long-list condition and the other for the Time 2 long-list condition. From these larger lists, 10-item word lists were created that appeared at Time 1 or Time 2. All words were presented equally often in the 60-item and 10-item list, and words were likewise counterbalanced for their appearance at Time 1 or Time 2.

Half of the participants first studied, made predictions, and were tested on a long-list of words (i.e., 60 words at Time 1) and then completed the same procedure a second time with a short list of words (i.e., 10 words at Time 2). The other half of the participants first studied, made predictions, and were tested on a short list of words (i.e., 10 words at Time 1) and then completed the same procedure a second time with a longer list of words (i.e., 60 words at Time 2). All other aspects of the procedure were identical to that used in Experiment 1.

**Results**

**JOL sensitivity.** Mean JOLs and the mean percentage of items recalled in Experiment 4 are presented in Figure 4. JOLs were analysed in a $2 \times 2$ (Time: 1, 2) × 2 (List Length: 10, 60) repeated-measures ANOVA. Results showed that JOLs for the 10-item condition ($M = 52.65$, $SD = 18.19$) were reliably greater than JOLs for the 60-item condition ($M = 36.46$, $SD = 14.15$), $F(1, 56) = 14.22$, $\eta_p^2 = .22$. The main effect of Time was not reliable, $F < 1$. Finally, a marginal Time × List Length interaction was evident, $F(1, 56) = 2.97$, $p = .09$, $\eta_p^2 = .05$. Specifically, while JOLs did not differ between the 10-item and 60-item conditions at Time 1, $t(26) = 1.66$, $p = .11$, JOLs were reliably higher for the 10-item condition compared with the 60-item condition at Time 2, $t(26) = 3.49$, Cohen’s $d = 1.37$.

**Calibration.** Predicted and actual recall performance were analysed in a $2 \times 2$ (Time: Time 1, Time 2) × 2 (List Length: 10 words, 60 words) × 2 (Measure: JOL, recall) mixed-factor ANOVA.
There was no main effect of Measure or Time ($F_{s} < 1$) but a main effect of Condition, $F(1, 52) = 72.76$, $\eta_{p}^2 = .58$, was evident, as was a significant Measure $\times$ Condition interaction, $F(1, 52) = 16.66$, $\eta_{p}^2 = .24$. More importantly, a reliable three-way interaction of Time, Measure, and Condition was evident, $F(1, 52) = 6.35$, $\eta_{p}^2 = .11$.

Critically, the three-way interaction demonstrates a change in calibration between Time 1 and Time 2. In particular, at Time 1, JOLs were significantly lower than recall for the 10-item condition, $t(13) = 3.05$, Cohen's $d = 1.15$, whereas JOLs were significantly higher than recall for the 60-item condition, $t(13) = 4.08$, Cohen's $d = 1.47$. It is likely that this under-confidence for the 10-item condition at Time 1 is anomalous given several prior experiments demonstrating strong calibration after studying a list of 10 items. An entirely different pattern of data was evident at Time 2. Specifically, no significant differences were found between JOLs and recall for either the 10-item ($t < 1$) or 60-item list length conditions, $t(13) = 1.91$, $p = .18$. Thus, participants can become well calibrated for a long list if provided with the comparative information gleaned from experience with a prior list.

Item position effects. As in previous experiments, we examined calibration as a function of item position for the long-list condition in a 2 (Measure: JOL, recall) $\times$ 2 (List Position: first 10, last 10) $\times$ 2 (Time: 1, 2) mixed-factor ANOVA (see Table 1). There was a marginal main effect of Measure, $F(1, 26) = 3.09$, $p = .09$, $\eta_{p}^2 = .11$, while the main effects of Time ($F < 1$) and List Position were not reliable, $F < 1$. However, Measure reliably interacted with Time, $F(1, 26) = 5.47$, $\eta_{p}^2 = .17$. Follow-up tests demonstrated that Time 1 JOLs exceeded recall, $t(13) = 3.16$, Cohen's $d = 1.75$, while no difference was found between JOLs and recall at Time 2, $t < 1$. Thus, participants were well calibrated at Time 2, but were overconfident at Time 1. No other interactions were evident.

To explore metacognitive sensitivity, a 2 (List Position: first 10, last 10) $\times$ 2 (Time: 1, 2) mixed-factor ANOVA was conducted. This analysis revealed no reliable effects of List Position ($F < 1$) or Time, $F(1, 26) = 1.17$, $p = .29$, $\eta_{p}^2 = .14$, nor did List Position interact with Time, $F(1, 26) = 1.13$, $p = .30$, $\eta_{p}^2 = .04$. Thus, there were no reliable differences in JOLs from the beginning to the end of the list based on the study-test trial (i.e., either Time 1 or Time 2).

**GENERAL DISCUSSION**

In the experiments reported, participants studied and made predictions for a long or short list of words, yielding several notable findings. Across all four experiments, participants demonstrated metacognitive sensitivity to the amount of material to be remembered (TBR). That is, participants provided lower JOLs for longer lists of words (i.e., either 100 or 60 words) compared with shorter lists of words (i.e., 10 words). However, calibration appears to suffer when participants must learn a long list of words, in contrast to the generally excellent calibration evident when participants learned a short list of words. Specifically, when lists of 60 or 100 items were studied, participants exhibited substantial overconfidence whereas, with shorter lists (10 or 30 items), participants’ predictions were well calibrated with later memory performance. Additionally, calibration did not improve for long lists when participants were given information about both list length conditions (Experiment 2) or specifically warned (Experiment 3) that a long list of words is difficult to remember (cf. Rhodes & Castel, 2008a). Only when participants experienced learning both long and short lists of words did memory predictions adequately account for the amount of material TBR (Experiment 4). These data indicate that while theory-based knowledge may not be sufficient for well-calibrated predictions with large amounts of material TBR, prior experience resulted in better-calibrated predictions.

Data reported for each of experiments focused on measures of absolute accuracy. How does list length impact relative accuracy? While our primary interest was in calibration, for comprehensiveness we report Kruskal-Goodman gamma correlations (Nelson, 1984) for all experiments in the Appendix. Conceptually, resolution captures the degree to which predictions distinguish between items that will and will not be remembered. Positive gamma correlations result when recalled words are given high JOLs and low JOLs are provided for words that are not remembered; in contrast, negative gamma correlations result when high JOLs are provided for words that are not remembered and recalled words are given low JOLs. Data for Experiments 1, 2, and 3 indicate that gamma correlations were numerically but not reliably different between the short and long list length conditions. This pattern also held for
Experiment 4 at Time 1 and Time 2. Thus, relative accuracy was not greatly influenced by the amount of material TBR, as gamma correlations did not vary between conditions for any of the 4 experiments.

The present experiments further suggest that the sensitivity of JOLs is influenced by the amount of material TBR, consistent with a direct access account of metacognition (Arbuckle & Cuddy, 1969). That is, long lists should serve to decrease memory strength and yield low JOLs, while memory strength should be greater for participants in the short list condition resulting in higher JOLs. Critically, while data for metacognitive sensitivity accorded with a direct access approach, the calibration data yield a differing pattern. Specifically, the direct access approach has difficulty accommodating the high levels of overconfidence evident in the long-list condition.

Instead we suggest that such a discrepancy between predictions and performance can be accounted for by an inferential framework such as that proposed by Koriat (1997). In particular, the current data indicate that participants focus on characteristics of the items presented (an intrinsic cue) at the expense of other information about the study phase that is diagnostic of performance, such as the length of the list presented. Such findings are consistent with work (e.g., Koriat et al., 2004) demonstrating that participants’ predictions may fail to take into account important factors (e.g., retention interval) affecting memory performance and instead focus on item-specific characteristics (e.g., relatedness). Prior list-learning experience potentially allows participants to shift from a focus on item-specific information to other, more diagnostic information associated with the study context, yielding improved calibration.

The data reported in the current study are consistent with previous work demonstrating that JOLs are not always able to fully appreciate the impact of conditions which produce interference (e.g., Metcalfe et al., 1993; Rhodes & Castel, 2008b; but see Maki, 1999). In addition, results from Experiment 4 accord with those reported by Rhodes and Castel (2008b). They observed that participants’ predictions were only well calibrated for the interference engendered by part-set cues when given prior experience with a similar task. In a similar manner, Experiment 4 showed that participants were well calibrated for a long list of words only when provided with a prior list-learning experience. Thus, participants may be best able to make predictions under conditions of interference when they have some prior experience with the task compared with relying on theory-based knowledge. While the focus of Experiment 4 was on prior experience as defined by a previous study-test trial, it remains unclear whether prior study or prior test alone are sufficient to promote a shift from theory to experience-based knowledge for the amount of material TBR.

Summary and conclusions

In summary, the current study is the first, to our knowledge, to examine predictions of memory performance as a function of the amount of information TBR. Results consistently demonstrated a dissociation between metacognitive sensitivity and calibration, consistent with prior work (e.g., Moulin et al., 2000). Specifically, while participants’ predictions were sensitive of list length, predictions of memory performance were poorly calibrated when participants were asked to learn a long list of words. Indeed, participants were only well calibrated for a long list of words when they were provided with experience on a list-learning task. It is perhaps surprising that participants need experience in order for predictions to be well calibrated with the amount of material TBR, and suggests that future research would be well served by further examining other potential methods to aid metacognitive awareness of large amounts of material TBR. Students are frequently required to learn various amounts of material in preparation for exams. The present data suggest that students may not adequately account for the amount of material to be covered on their next exam when studying.

REFERENCES


is misleading as a metamnemonic index. *Journal of Experimental Psychology: General, 127*, 55–68.


Resolution data and the pattern of results for Experiments 1–4 are provided below in Table A1. We measured resolution via Kruskal-Goodman gamma correlations (G).

<table>
<thead>
<tr>
<th>Experiment</th>
<th>Short-list condition</th>
<th>Long-list condition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Experiment 1</td>
<td>.37 (.72)*</td>
<td>.37 (.16)*</td>
</tr>
<tr>
<td>Experiment 2</td>
<td>.14 (.65)</td>
<td>.40 (.28)*</td>
</tr>
<tr>
<td>Experiment 3</td>
<td>.20 (.69)</td>
<td>.46 (.32)*</td>
</tr>
<tr>
<td>Experiment 4, Time 1</td>
<td>.09 (.65)</td>
<td>.40 (.30)*</td>
</tr>
<tr>
<td>Experiment 4, Time 2</td>
<td>.25 (.64)</td>
<td>.36 (.26)*</td>
</tr>
</tbody>
</table>

Standard deviations provided in parentheses. For all experiments, 10-item word lists were employed for the short-list condition. For the long-list condition, Experiments 1–3 employed 100-item word lists and Experiment 4 employed 60-item word lists. *G significantly greater than chance (zero). For all experiments, the magnitude of the correlations did not differ between the long- and short-list conditions, ps > .05.