Mem Cogn (2015) 43:910–921
DOI 10.3758/s13421-015-0516-4

When will bigger be (recalled) better? The influence of category size on JOLs depends on test format

Kathleen L. Hourihan · Jonathan G. Tullis

Published online: 11 March 2015

Abstract Although it is well known that organized lists of words (e.g., categories) are recalled better than unrelated lists, little research has examined whether participants can predict how categorical relatedness influences recall. In two experiments, participants studied lists of words that included items from big categories (12 items), small categories (4 items), and unrelated items, and provided immediate JOLs. In Experiment 1, free recall was highest for items from large categories and lowest for unrelated items. Importantly, participants were sensitive to the effects of category size on recall, with JOLs to items from big categories actually increasing over the study list. In Experiment 2, one group of participants was cued to recall all exemplars from the categories in a blocked manner, whereas the other group was cued in a random order. As expected, the random group did not show the recall benefit for big categories over small categories observed in free recall, while the blocked group did. Critically, the pattern of metacognitive judgments closely matched actual cued recall performance. Participants' JOLs were sensitive to the interaction between category size and output order, demonstrating a relatively sophisticated strategy that incorporates the interaction of multiple extrinsic cues in predicting recall.

Keywords Metacognition · Recall · Relatedness · JOLs

K. L. Hourihan (✉)
Department of Psychology, Memorial University of Newfoundland, St. John's, NL A1B 3X9, Canada
e-mail: khourihan@mun.ca

J. G. Tullis
Department of Educational Psychology, University of Arizona, Tucson, AZ, USA

To-be-remembered information often exists in categorized lists. For instance, doctors may need to remember the possible symptoms of a disease, shoppers may need to remember the baking supplies that they need to buy, and students learning German may need to remember translations of words for days of the week. Accurately predicting whether this categorized information will be remembered is important because it affects the choices we make about our memories (Nelson & Narens, 1990, 1994). For example, memory judgments guide what we choose to study (Thiede & Dunlosky, 1999), how we choose to study (Karpicke, 2009; Tullis, Benjamin, & Fiechter, 2015), and how long we choose to study (Nelson & Leonesio, 1988; Tullis, Benjamin, & Liu, 2014). Furthermore, the accuracy of these memory predictions largely determines memory performance (Thiede, Anderson, & Therriault, 2003; Tullis & Benjamin, 2011) and may be more consequential than individual differences in memory ability (Benjamin, 2008). In the current study, we explore metacognitive monitoring during study of a list of related items and demonstrate that learners are able to predict the effects that categorical relatedness has on future recall performance.

Relatedness and memory

There has been a recent renewed interest in analyzing the impact of relations among stimuli rather than analyzing stimuli as discrete, independent events. For example, the *relations* among individual events have recently been proposed to influence interpretation of ambiguous events (Ross & Bradshaw, 1994; Tullis, Braverman, Ross, & Benjamin, 2014), the transfer of knowledge to novel problems (Ross & Kennedy, 1990), categorization (Ross, Perkins, & Tenpenny, 1990; Wahlheim, Dunlosky, & Jacoby, 2011), the spacing effect in memory (Benjamin & Tullis, 2010), and even memory for individual instances in related pairs (MacLeod,

Pottruff, Forrin, & Masson, 2012; Tullis, Benjamin, & Ross, 2014; Wahlheim & Jacoby, 2013). The impact of relations among studied items in higher cognition and memory has been researched extensively; yet, the impact of relations among studied items on metamemory remains largely unexplored. While research has typically focused on the cognitive implications of relations among studied items, here we examine how relations influence metacognitive monitoring.

Many studies have demonstrated that related items in a list are remembered better than unrelated items (for a review, see Kausler, 1974, pp. 345–390). Not only are related items better remembered than unrelated items but also items from large categories are recalled better than items from small categories. That is, free recall of lists of equivalent length increases as the size of categories included increases (Tulving & Pearlstone, 1966). In Tulving and Pearlstone's study, recall was poorest for lists composed of categories of size 1 (i.e., lists of unrelated items), better when composed of categories of size 2 (i.e., associated pairs), and best when composed of categories of size 4.

Various mechanisms have been suggested to explain the mnemonic benefits of relatedness. According to reminding theories, later presentations remind learners of earlier related items and memory for the involved instances is enhanced (Hintzman, 2010; Tullis et al. 2014a, b). Similarly, Rundus (1971) suggested that the increased and strategic rehearsal of earlier category members during later presentations of other category members leads to better memory performance for related over unrelated items. Alternatively, Hunt and Seta (1984) proposed that successful recall depends on encoding both item-specific (i.e., distinctive aspects of individual items) and relational (i.e., similarities among items in the set) information. Categories, especially large categories, include many items that are similar to other studied items; learners notice similarities at the time of encoding and organize the items based around those categories in a way that supports later retrieval (Einstein & Hunt, 1980; Hunt & Einstein, 1981). Evidence shows that learners tend to organize recall output order based on categorical relations present in the study list, even when those items were presented in a randomized order at study (Bousfield, 1953). This indicates that participants notice categorical relations that occur across a study list, and use that categorical information to organize rehearsal and subsequent recall.

Monitoring relatedness

While many studies have examined the mnemonic consequences of relatedness, only one study has specifically shown that categorical relations among studied items also increase mnemonic predictions. In Matvey, Dunlosky, and Schwartz (2006), participants studied lists of words that contained blocks of four related items and blocks of four unrelated items, and made item-by-item judgments of learning (JOLs). Results showed that participants were able to accurately predict that related items would be recalled better than unrelated items, although relatedness affected recall much more than it affected JOLs. Detailed analyses of the JOLs showed that judgments were sensitive to the changes in relatedness experienced during the study list. That is, because blocks of related and unrelated items were ordered randomly during the study list, participants did not know whether a given block of items would be related until at least the second word in a block was presented. Accordingly, JOLs for the second, third, and fourth items in a block of four related words were higher than for the first item in that block; JOLs for the four items in unrelated blocks did not differ from one another. Moreover, the differences in JOLs between unrelated and related blocks was not significant when considering only the first blocks of each type that participants experienced: JOLs did not differentiate between related and unrelated blocks of items until participants had encountered at least one block of each type. This indicates that participants were able to keep track of the overall composition of the study list, and to update their JOLs based on changes in extrinsic factors experienced over the course of the study list. We extended this study by analyzing whether learners account for category size in their judgments.

Because JOLs are the product of an inferential process that combines information from intrinsic cues (information about the item itself, e.g., word frequency), extrinsic cues (information about the study context, e.g., repetition), and mnemonic cues (information about subjective memorability, e.g., ease of retrieval), learners may predict better recall for category members for several different reasons (Koriat, 1997). First, learners may use the nonanalytic (i.e., nonconscious) cue of fluency of processing to judge memorability (Begg, Duft, Lalonde, Melnick, & Sanvito, 1989; Benjamin, Bjork, & Schwartz, 1998; Hertzog, Dunlosky, Robinson, & Kidder, 2003; Matvey, Dunlosky, & Guttentag, 2001). Fluency of processing is a mnemonic cue that can have a large, nonconscious influence on monitoring (Koriat, 1997). Earlier presentations of category members may make the processing of later category members more fluent, and, consequently, learners may judge category members as increasingly memorable throughout a category list. This view suggests that JOLs should increase continuously across a category, which was not found in Matvey et al. (2006).

Alternatively, as proposed by Matvey et al. (2006), learners may use controlled, analytic inferences about category membership when assigning JOLs to categorized items. Learners may rate categorized items as more memorable than unrelated items because the category label provides a retrieval mediator that facilitates successful retrieval during testing (Hertzog & Dunlosky, 2004). Learners may recognize that they have a retrieval cue for stimuli (i.e., the category label) and therefore rate words associated with that retrieval cue as more

memorable. In this view, learners may consider category membership a dichotomous variable and rate categorized items as more memorable than unrelated items, regardless of category size and position within a category.

Finally, we suggest that learners rate categorized items as more memorable because they consciously rehearse and strategically plan to output those items together. Learners group categorized instances together during encoding, rehearse them together, and use these groupings as an effective output strategy (Rundus, 1971). Therefore, because they recognize their output strategy for these items is effective, learners rate categorized instances as more memorable than unrelated items. Rather than focusing on the connection of the items to the category label as in the mediator hypothesis, this view suggests that the relationships among the related items drives learners' rehearsal and mnemonic predictions. Learners may consider the expected test format (and its match with how they are encoding and rehearsing the stimuli) when assigning JOLs and give ratings based upon how well their output strategy aligns with the test format. Prior research supports this idea, as expected test format has an impact on student's mnemonic predictions (and consequently their study choices; Finley & Benjamin, 2012; Thiede, 1996; Thiede, Wiley, & Griffin, 2011). We investigated these three different explanations across two experiments.

The goal of the present study was to examine whether learners can accurately predict how categorical relatedness and category size will affect future recall. Participants studied lists of words that contained unrelated items (one-item categories), small categories (with each category containing four exemplars in total), and big categories (with each category containing 12 exemplars in total). Items of all three types (big categories, small categories, one-item categories) were not presented in blocks (cf. Matvey et al., 2006), such that category membership status for a given item was not necessarily obvious until at least partway through the study list. Participants provided JOLs immediately following each item in the list and then completed a free recall test. Although blocking category exemplars at study usually leads to better recall than spacing them over the list (e.g., Dallett, 1964; but see Borges & Mandler, 1972) categorized items were nevertheless expected to be recalled better than unrelated items, and big categories were expected to be recalled better than small categories. The goal of Experiment 1 was to determine whether participants could accurately predict how category size would affect free recall performance. If learners are using category membership as a dichotomous, analytic cue, we do not expect to see differences in JOLs between big and small categories (although both should be rated higher than unrelated items). In Experiment 2, we varied how memory was tested to determine whether participants' JOLs account for how recall of categorized information can be disrupted by altering output order.

Experiment 1

In the first experiment, our goal was to examine whether participants could predict how category size would influence their free recall performance. Participants made immediate JOLs for study lists containing items from big categories (12 items), items from small categories (four items; three small categories were included in each list), and items from one-item categories (12 items each selected from different categories). Based on past research on how category size influences recall (Hunt & Seta, 1984; Tulving & Pearlstone, 1966), it was expected that words from big categories would be recalled the best, followed by words from small categories, with words from one-item categories recalled the worst. Moreover, we expected participants to display category clustering at recall as is typically seen with categorized items (e.g., Einstein & Hunt, 1980; Hunt & Einstein, 1981; Hunt & Seta, 1984; Tulving & Pearlstone, 1966), even when the items are not blocked at study (e.g., Bousfield, 1953; Dallett, 1964). The critical question was whether JOLs made at the time of encoding would successfully account for the benefits of relational processing (an extrinsic cue) on subsequent recall. Although we predicted that participants would provide higher JOLs to both types of categorized items than to unrelated items (as seen in Matvey et al., 2006), it was not clear whether participants would be able to accurately predict the effects of *category size* on free recall.

Method

Participants

Thirty-seven introductory psychology students at Indiana University participated for partial course credit.

Materials

Sixteen categories, including *units of time*, *relatives*, *furniture*, and *vegetables* were selected from the Van Overschelde, Rawson, and Dunlosky (2004) norms. The 12 most frequent responses in each selected category were used as potential targets. Additionally, one exemplar from each of 48 different unused categories was compiled into the one-item categories condition. Categories were randomly assigned to big or small conditions. During each study list, participants studied 12 exemplars from one category (the big category), four exemplars from each of three different categories (the small categories), and 12 exemplars from the filler list (the one-item categories). For the small categories, four of the 12 exemplars from a category were randomly selected to be studied.

Each study list was divided into thirds, such that each third had four exemplars from the big category, four exemplars

from one of the small categories, and four one-item filler exemplars. The order of exemplars within each third of the list was randomized. All exemplars of one small category were presented before another small category was introduced; in this manner, the average spacing between exemplars from the big category was the same as the average spacing between exemplars from small categories. Participants were yoked to each other in groups of three in order to counterbalance the serial position of the big, small, and one-item exemplars in each list. Where the first participant saw an exemplar from a big category, the second participant saw an exemplar from a small category and the third participant saw an exemplar from the one-item category.

Procedure

The experiment was programmed using the Psychophysics Toolbox extensions in MATLAB (Brainard, 1997) and utilized a three-condition (big categories, small categories, and one-item categories) within-subjects design. Participants read the instructions on individual computers in 10 different testing booths. The instructions asked participants to remember the studied words for a later memory test and to rate their likelihood of remembering each studied word on a scale of 1 to 4, where 1 was labeled "*will not remember,*" 2 was labeled "*probably will not remember,*" 3 was labeled "*probably will remember,*" and 4 was labeled "*will remember.*" Participants were not informed about the presence of categorical relations among items in the study list, nor were they informed about the exact nature of the upcoming memory test. An individual target was presented in the middle of the computer screen for 2 seconds in Arial 50-point font before the rating scale also appeared on the bottom of the screen. Participants typed their JOL, the screen was cleared, and the next target appeared in the center of the screen. They were not informed of the list length or current trial number during study. Participants studied and rated 36 words before the free recall test began. During the free recall test, participants were instructed to type any words that they could remember from the immediately prior list. Participants pressed the question mark button to end the test whenever they could remember no more targets. They completed four study–test cycles.

Results

First, we analyzed whether study list interacted with mean recall or JOLs. A two-way repeated measures ANOVA on recall revealed no evidence that study list interacted with category size, $F(6, 216) = 1.07$, $p = .38$, $\eta_p^2 = .03$, or had a main effect, $F(3, 108) = 1.47$, $p = .29$, $\eta_p^2 = .04$. Similarly, a two-way repeated measures ANOVA on JOLs showed no evidence that list interacted with category size, $F(6, 216) = 0.42$, $p = .87$, $\eta_p^2 = .01$, or had a main effect, $F(3,108) = 2.02$, $p = .12$, $\eta_p^2 = .05$. Therefore, in these analyses, we collapsed recall and JOLs across study list. Analyses focus on planned orthogonal contrasts, with the first contrast comparing the unrelated items to the mean of the two types of categorized items, and the second contrast comparing the two types of categorized items to one another (i.e., small categories vs. big categories).

Recall

Proportion correct recall is displayed in Fig. 1. A one-way repeated measures ANOVA on recall revealed a significant effect of category condition, $F(2, 72) = 137.18$, $p < .001$, $\eta_p^2 = .79$. The first planned contrast revealed that participants recalled fewer words from the one-item categories, with the contrast between the one-item categories and the categorized items (i.e., big and small) being significant, $F(1, 36) = 255.77$, $p < .001$, $\eta_p^2 = .88$. Importantly, the second contrast between big and small categories was also significant, $F(1, 36) = 17.44$, $p < .001$, $\eta_p^2 = .33$, showing that, as predicted, participants recalled more from big categories than small categories.

We investigated how participants output items during free recall. For each participant, we calculated the adjusted ratio of clustering (ARC; Gerjuoy & Spitz, 1966; Roenker, Thompson, & Brown, 1971) within each list. The mean ARC was significantly greater than zero ($M = 0.61$, $SD = 0.18$), $t(36) = 20.78$, $p < .001$, $d = 3.42$, which indicates that participants output clusters of items from the same category during recall. Furthermore, we computed the average output position for items within each category size. The average output position for items from big categories was earlier ($M = 8.11$, $SD = 2.21$) than the average output position for items from both small categories ($M = 9.82$, $SD = 2.30$), $t(36) = 3.78$, $p < .001$, $d = 0.63$, and one-item categories ($M = 10.05$, $SD = 3.40$), $t(36) = 3.66$, $p < .001$, $d = 0.61$. Finally, we analyzed the proportion of categories participants accessed, which we defined as recalling at least one exemplar from a category. Astonishingly, every participant recalled at least one exemplar from each of the big categories that they studied. Participants accessed only ¾ of the small categories ($M = 0.75$, $SD = 0.14$).

Judgments of learning

Mean JOLs are displayed in Fig. 1. A one-way repeated measures ANOVA on JOLs revealed a significant effect of category condition, $F(2, 72) = 36.93$, $p < .001$, $\eta_p^2 = .50$. The first planned contrast shows that participants predicted lower recall for words from the one-item categories, relative to the categorized words, $F(1, 36) = 42.87$, $p < .001$, $\eta_p^2 = .54$. Critically, the second contrast between big and small categories was also significant, $F(1, 36) = 19.49$, $p < .001$, $\eta_p^2 = .35$, indicating

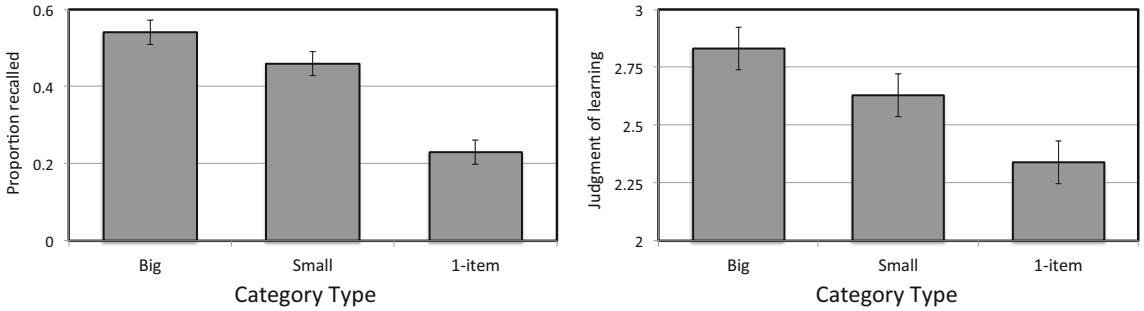

Fig. 1 Proportion recalled (left pane) and JOLs (right pane) as a function of category type in Experiment 1. Error bars show the 95% within-subjects confidence intervals (Loftus & Masson, 1994)

that participants predicted that they would recall more from big categories than from small categories.

Mean JOLs as a function of study position within each category are displayed in Fig. 2. For each participant, we computed the correlation between an item's study position within each category condition and the item's JOL. For big categories, the correlation between an item's study position within a category and its JOL was positive ($r = 0.16$) indicating that JOLs increased across the members of the big category. For one-item categories, the correlation was slightly negative ($r = -.04$) and smaller than that from big categories, $t(36) = 2.43$, $p = .02$, $d = 0.41$. For small categories, we computed the correlation between JOL and study position within each of the three small categories in each list, rather than considering the three sequentially presented small categories as a single condition. Within each small category the correlation between study position and JOL was positive ($r = .14$) and not different than that found in the big category, $t(36) = .76$, $p = .31$). This indicates that within each big and small category, the JOLs increased with the position of the item within the category. When considering the whole small condition (i.e., including all three small categories presented across the study list), a decline in JOLs was observed when new small categories were introduced, as shown at serial positions 5 and 9 in Fig. 2.

Discussion

In our first experiment, the goal was to determine whether participants would appropriately account for the influence of category size on recall when making immediate JOLs. Category size influenced recall in the manner expected: Big categories were recalled better than small categories, which were recalled better than unrelated items. Importantly, JOLs accurately corresponded to the observed pattern of recall: JOLs were highest for big categories, slightly lower for small categories, and lowest for unrelated items. Our analyses of the JOLs show that participants were very sensitive to the presence of categorical information during the study list, and the effect of category size was as large on JOLs as it was on recall. JOLs generally increased as exemplars were added to a category, resulting in the observed positive correlation between serial position and JOL for the big categories. A similar pattern was observed when considering each individual small category (cf. Matvey et al., 2006). JOLs for one-item categories (unrelated items), however, decreased over the study list,

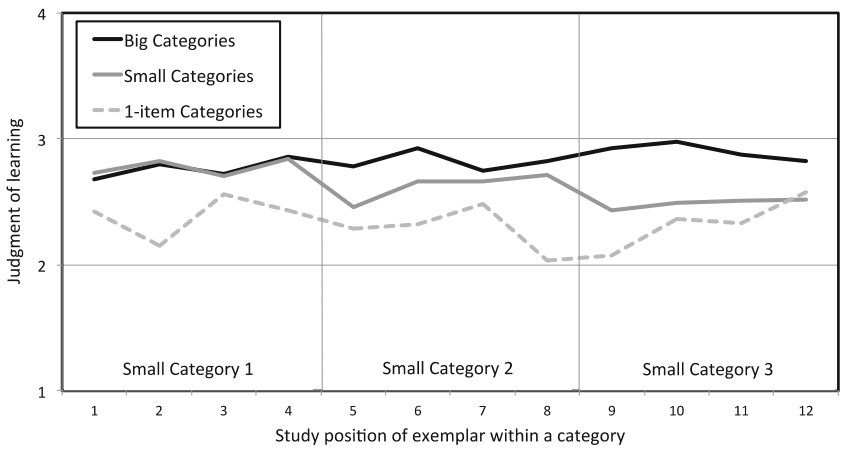

Fig. 2 Mean judgments of learning (JOLs) as a function of study position within each category in Experiment 1

as has been observed by others (e.g., Matvey et al., 2006; Tauber & Dunlosky, 2012). Thus, our results show that learners are able to accurately predict the effects of categorical relatedness on recall, even when category members are spaced across the study list. Moreover, they accurately predicted the effect of category size on free recall because they slightly increased recall predictions across members of a category. These results seem most consistent with the nonanalytic fluency account of assigning JOLs to related items, as JOLs increased across position with a category.

We also observed significant clustering of categories during free recall. This was unsurprising, given the past 60+ years of research on recall of categorical information (e.g., Bousfield, 1953; Dallett, 1964; Einstein & Hunt, 1980; Hunt & Einstein, 1981; Hunt & Seta, 1984; Tulving & Pearlstone, 1966). Categorically related items may be recalled better than unrelated items because they are encoded in a relational manner (Hunt & Seta, 1984). The relational information serves to organize rehearsal and ultimately output order; items that are related are rehearsed together and are later recalled together (cf. Rundus, 1971). Are participants aware that the benefits of categorical information in memory are dependent on their ability to output items in clusters? In other words, is the influence of category relatedness due solely to nonanalytic influences of fluency or do learners deliberately assign JOLs based upon expected test conditions? We explored this question in our second experiment.

Experiment 2

Thus far, we have considered how categories of different sizes affect JOLs and memory performance under conditions of free recall, in which participants have considerable control over how they choose to group items in rehearsal and at output. We now consider whether participants are able to predict how category size affects memory performance under different testing conditions: cued recall. Considering categorized study lists in particular, research has shown that categories of different sizes are affected differently by category-label cues at recall (e.g., Tulving & Pearlstone, 1966). Tulving and Pearlstone's study examined recall of items from categories of different sizes and compared groups who recalled freely with those who were cued with the category labels at the time of test. Their results showed that category-label cuing at recall benefited smaller categories more than larger categories. Groups tested under cued recall conditions recalled more *categories* than those tested under free recall conditions, but the groups did not differ substantially in the number of *items per category* recalled. That is, the benefits of category cuing at test primarily seem to be in terms of alerting participants that a given category was presented at all, not necessarily of which items belonging to that category had been presented. Because large categories are less likely to be forgotten entirely than small categories, small categories benefit more from category cues at test than large categories (Tulving & Pearlstone, 1966).

In our second experiment, we examined whether participants would alter their mnemonic predictions based upon the differential effects of cuing on memory for items from different-sized categories. The study phase was identical to the first experiment: Participants studied lists of categorized target words and provided immediate JOLs. The test was changed from free recall to cued recall. On each test trial, participants were given a category label (e.g., "furniture") and asked to recall a studied exemplar from that category (e.g., "couch"). Critically, participants were assigned to one of two groups that differed on the order of the cues they received at test. In the random test order condition, the category labels were presented in an entirely random order. In the blocked test order condition, however, all of the same category labels were presented back-to-back. The blocked test condition was intended to cue recall in a manner similar to the way in which participants would choose to output items if given a free recall test: in clusters of categories. The random test condition was designed to cue recall in a manner that would be dissimilar to how participants rehearse related information (Rundus, 1971) and would likely disrupt how participants would otherwise output items. Prior to beginning the cued-recall test, participants would likely plan to output the studied words based on the categorical structure of the list (i.e., a plan to cluster categorically related words at recall, as observed in Experiment 1) because learners' predictions about an unspecified memory test often align closely with the demands of free recall tests (e.g., Benjamin, 2003; Tullis & Benjamin, 2011). Cuing participants in a random order should disrupt their retrieval plan, especially for larger categories, and reduce overall recall performance (similar to part-list cuing effects; cf. Slamecka, 1968).

For the one-item categories, however, we expect performance in Experiment 2 to increase relative to Experiment 1. As described above, the benefit of category cuing (relative to free recall) is primarily in terms of increasing the number of categories recalled, rather than influencing the number of items per category (Tulving & Pearlstone, 1966). In our study, the unrelated items are effectively 12 different categories of size 1; each category label therefore cues one studied item uniquely. Category cuing should therefore improve recall of items from the one-item categories (relative to performance under free recall conditions in Experiment 1) in both testing conditions.

We predicted that the blocked recall condition should look much like in Experiment 1, but with a boost to recall for the one-item categories in particular. In the random test order group, recall of the big categories in particular should be disrupted the most. The inability to output items from the same category consecutively should disrupt participants'

retrieval plans, causing them to forget more items from that category, and likely increase repetition errors. We expected JOLs in the blocked test order group to reflect the effects of category size, as in Experiment 1. The important question is whether participants in the random test order group will recognize that the cued recall test does not allow them to organize memory at output in the way in which they would like to (clustered by category), and reduce their JOLs to big category items accordingly.

Method

Participants

Ninety-three introductory psychology students at Indiana University participated for partial course credit. Participants were alternatively assigned to the random (n = 47) and the blocked (n = 46) test order conditions.

Materials

The same categories that were used in Experiment 1 were used in Experiment 2, but the specific exemplars chosen from each category differed. The specific exemplars were less frequent responses to each selected category than in Experiment 1, in order to reduce the possibility of guessing during the cued recall test (e.g., Smith, Ward, Tindell, Sifonis, & Wilkenfeld, 2000).

Procedure

The study procedure in Experiment 2 was identical to that in Experiment 1, but the testing procedure differed. During the test phase, participants were presented with the category label and were asked to type in a studied exemplar from that category. In the random test order condition, the order of the category labels was randomized. In the blocked test order condition, category labels were presented in a more structured order, as shown in Fig. 3. Six category labels from the one-item categories were presented first. Then, the big and small category labels were blocked, such that the big category label was repeated 12 times in a row and each small category label was repeated 4 times in a row. For a random half of the study lists, the three small category labels were presented before the big category label, and for the other half of the study lists, the big category label was presented before the small category labels. Each small category label was presented blocked together before the next small category label was presented (e.g., "furniture" "furniture" "furniture" "furniture" "relative"). At the end of each test list, the six untested category labels from the one-item categories were presented. Structuring the test output in this order resulted in equal average test positions for each category across random and blocked test order conditions.

Results

We do not expect to see differences in List 1 JOLs between test conditions because learners across conditions had no experience with differentiated tests at that point. However, learners experienced the test conditions during the first test phase and may have changed their predictions and study strategies at that point. Therefore, we analyzed the results for List 1 separately from Lists 2 to 4 (which we averaged together). Each of these datasets were analyzed using orthogonal contrasts on the category size variable and including test type as a between-subjects factor. As in Experiment 1, the first contrast compared one-item category words to the average of the two categorized (i.e., big and small) word conditions and the second contrast compared big and small categories.

Recall

Mean proportion correct recall is presented in Fig. 4. First, the main effect of group was not significant, $F(1, 91) = 2.37$, $p = .127$, $\eta_p^2 = .03$. The first orthogonal contrast comparing recall of one-item category words to the mean of the categorized words was marginally significant, $F(1, 91) = 3.03$, $p = .085$, $\eta_p^2 = .03$, as was the second contrast comparing recall of big category words to small category words, $F(1, 91) = 3.50$, $p = .064$, $\eta_p^2 = .04$. Importantly, the interaction between group and the second contrast was significant, $F(1, 91), = 4.20$, $p = .043$, $\eta_p^2 = .04$, indicating that while participants in the blocked test group recalled more items from big categories than small categories, those in the random testing group did not.

A similar pattern of average recall results was found across Lists 2 to 4. The main effect of group was not significant, $F(1, 91) = 0.95$, $p = .332$, $\eta_p^2 = .01$. The first orthogonal contrast comparing recall of one-item category words to the mean of

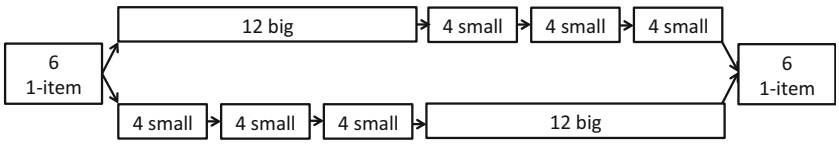

Fig. 3 The order of the tested exemplars for the blocked test order condition in Experiment 2. For each subject, the big category was presented before the small categories on half of the study lists (top row), and for the other half of the study lists, the small categories were presented before the big category (bottom row)

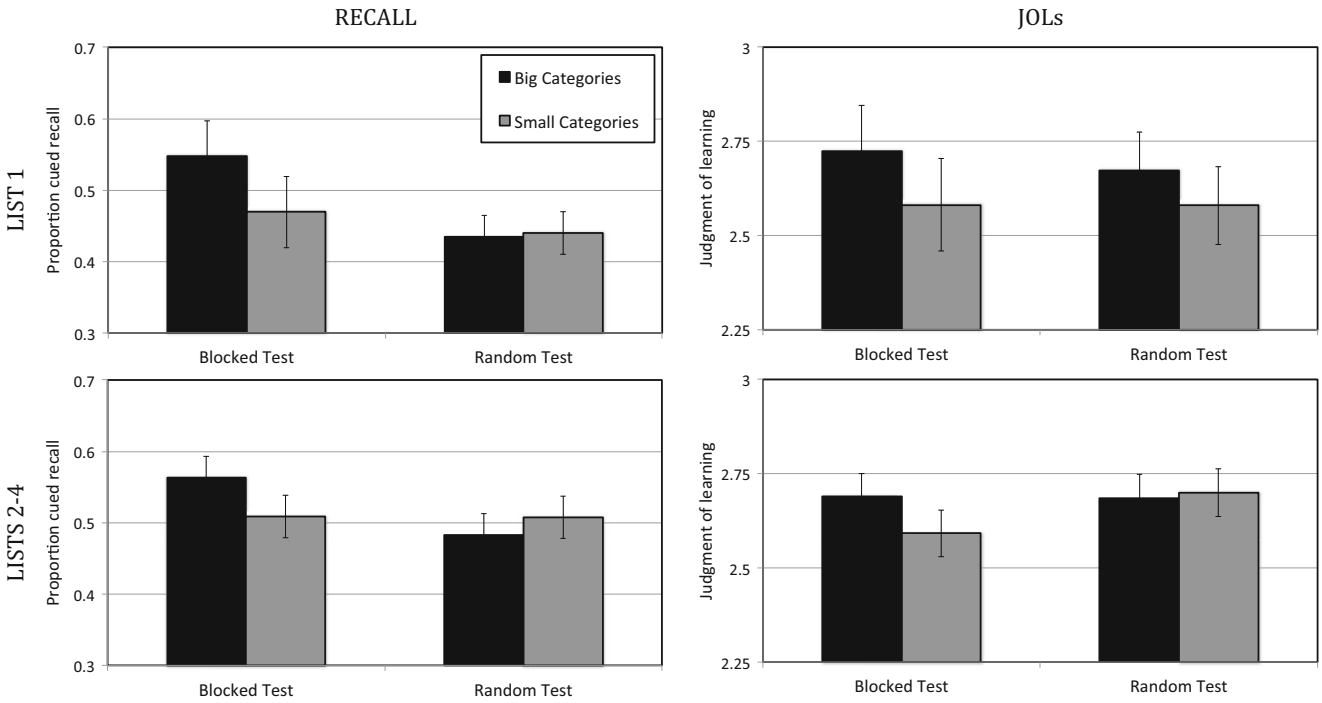

Fig. 4 Proportion cued recall (left column) and JOLs (right column) from List 1 (top row) and Lists 2 to 4 (bottom row) in Experiment 2. Error bars show the within subjects 95% confidence intervals (Loftus & Masson, 1994)

the categorized words was not significant, $F(1, 91) = 0.20$, $p = .656$, $\eta_p^2 = .002$, nor was the second contrast comparing recall of big category words to small category words, $F(1, 91) = 2.23$, $p = .134$, $\eta_p^2 = .03$. More importantly, however, the interaction between group and the second contrast was significant, $F(1, 91), = 7.32$, $p = .008$, $\eta_p^2 = .07$, indicating that while participants in the blocked test group recalled more items from big categories than small categories, those in the random testing group did not. Thus, the pattern of recall in Lists 2 to 4 matched the pattern of recall in List 1.

Unlike the previous experiment, recall of the one-item category words was not significantly lower than recall of exemplars from big or small categories, as evidenced by the first contrasts in the above analyses. In fact, in the blocked test condition, recall of the one-item category words ($M = 0.53$, $SD = 0.19$) was numerically greater than recall from the small categories ($M = 0.50$, $SD = 0.19$). In the random test condition, recall of the one-item category words ($M = 0.50$, $SD = 0.18$) was numerically greater than both recall from the big categories ($M = 0.47$, $SD = 0.18$) and the small categories ($M = 0.48$, $SD = 0.17$).

The presence of category cues at recall also improved access to the smaller categories, compared to Experiment 1. In the blocked test condition, learners recalled at least one exemplar from 97% of the big categories ($SD = 0.10$) and from 86% of the small categories ($SD = 0.17$). In the random test condition, learners recalled at least one exemplar from 97% of the big categories ($SD = 0.15$) and from 87% of the small categories ($SD = 0.18$).

JOLs

Mean JOLs are displayed in Fig. 4. For List 1, the main effect of group was not significant, $F(1, 91) = 0.035$, $p = .851$, $\eta_p^2 < .001$. The first orthogonal contrast comparing JOLs for one-item category words to the mean of the categorized words was not significant, $F(1, 91) = 0.93$, $p = .339$, $\eta_p^2 = .01$, while the second contrast comparing JOLs for big category words to small category words was significant, $F(1, 91) = 5.34$, $p = .023$, $\eta_p^2 = .06$. Unlike in recall, the interaction between group and the second contrast was not significant, $F(1, 91) = 0.24$, $p = .628$, $\eta_p^2 = .003$. These results show that participants gave higher JOLs to items from big categories than small categories, regardless of test condition, when they had no knowledge of how their memories would be tested.

After participants experienced how their memories would be tested, however, test type influenced how category size impacted JOLs. The main effect of group was not significant, $F(1, 91) = 0.28$, $p = .596$, $\eta_p^2 = .003$. The first orthogonal contrast comparing JOLs for one-item category words to the mean of the categorized words was significant, $F(1, 91) = 42.62$, $p < .001$, $\eta_p^2 = .319$, but the second contrast comparing JOLs for big category words to small category words was not, $F(1, 91) = 2.55$, $p = .113$, $\eta_p^2 = .03$. Importantly, the interaction between group and the second contrast was significant,

$F(1, 91)$, = 4.95, p = .028, η_p^2 = .05, indicating that while participants in the blocked test group gave higher JOLs to items from big categories than small categories, those in the random testing group did not. This pattern is identical to that observed in actual cued recall.

Despite the improvement in cued recall for words from one-item categories (relative to free recall in Experiment 1), JOLs for unrelated items were relatively low across test conditions and lists. For List 1, participants rated unrelated items as only somewhat memorable, both in the blocked test condition (M = 2.55, SD = 0.52) and in the random test order condition (M = 2.65, SD = 0.50). This pattern persisted across Lists 2 to 4, where unrelated items were once again rated as least memorable in the blocked test order condition (M = 2.84, SD = 0.42) and in the random test order condition (M = 2.51, SD =0.47). As indicated by the contrast above, participants continued to believe that unrelated items would be recalled worse than categorized items, despite their recall performance showing no difference.

Recall repetitions

In each cued recall test, we expected that participants' output strategies would be disrupted and they would often output the same target multiple times. We expected greater disruption in the random test order condition than in the blocked test order condition. We calculated the percentage of output items that were repeated in the output list. Participants in the random test order condition repeated the same targets more often (M = 0.27, SD = 0.14) than participants in the blocked test order condition (M = 0.15, SD = 0.11); $t(91) = 4.53$, $p < .0001$, d = 0.99.

Discussion

We replicated the novel findings of our first experiment: With no experience, learners rated items from bigger categories as more memorable than items from smaller categories. As predicted, blocked cued recall enabled learners to recall more items from big categories than small categories, but the randomized cued recall order led to equal performance across category size. Critically, after learners in the random test condition experienced one test, they changed their predictions such that items from big and small categories were rated as equally memorable, perfectly paralleling actual cued recall performance. Learners in the blocked test order maintained their higher predictions for big categories over small categories, a pattern that also persisted in cued recall performance. We will discuss the implications of these results in the general discussion.

Although performance with the unrelated (one-item category) items is tangential to our focus on the influence of category size on JOLs and recall, we note that recall of items from one-item categories was twice as high in Experiment 2 (M = .52) than in Experiment 1 (M = .23). As noted above, this was entirely expected. The smaller the category, the greater the improvement in recall from category-label cueing (relative to free recall conditions; Tulving & Pearlstone, 1966). In the one-item categories, each category label specifically cued one studied item uniquely, and this cue specificity greatly improved recall of the unrelated items. Interestingly, learners still predicted their recall of these items would be significantly worse than of categorized instances (both big and small), even for Lists 2 to 4, after they had experienced the cued recall test, in which learners (in both test conditions) recalled as many items from one-item categories as from small categories. This suggests that learners do not base mnemonic predictions entirely upon the amount recalled across category conditions during the prior test and reflects a glaring mistake in learners' ability to update metacognitive monitoring based upon prior experience (as in Tullis, Finley, & Benjamin, 2013). We note, however, that the individual items used in our study were not fully rotated across all conditions; although specific categories were randomly assigned to either the big category or small category condition as described above, the items assigned to the one-item condition were selected from different categories altogether. One should therefore be cautious in making any direct comparisons of the categorized items with the unrelated items.

General discussion

The goal of the present study was to determine whether learners could predict the effects of category size on recall under varying test conditions. Across two experiments, learners correctly accounted for the effects of categorical relatedness on memory. Learners rated big categories as more memorable than small categories under test conditions that produced this outcome (free recall and blocked cued recall), but did not differentially rate big and small categories when memory performance between them was equivalent (randomized cued recall).

This reveals sophistication in learners' metacognitive predictions and begins to shed light on the reasons learners rate categorized items as more memorable. The results show that the mediator hypothesis (i.e., that learners rate items with a mediator – the category name – as more memorable than items without a mediator; Hertzog & Dunlosky, 2004) is unlikely, because learners gave increased ratings to each successive study item within a category and distinguished between big and small categories in their JOLs. The mediator hypothesis would predict equally high JOLs for all categorized items, regardless of category size or how many other instances had been presented. Second, processing fluency cannot be the sole explanation, as learners weighted category size based upon the

expected test format. If learners relied exclusively on fluency when making predictions, we would not have observed the differences in JOLs between the two test order conditions in Experiment 2. The most promising explanation is that learners explicitly notice relationships during encoding, intentionally rehearse related items together, strategically plan to output them together at test, and therefore rate them as more memorable. Support for the idea that learners use an explicit strategy of rating related items as more memorable than unrelated items comes from Mueller, Tauber, and Dunlosky's (2013) study on word pairs. Learners' predictions of categorical relatedness on memory may reflect a combination of all three types of cues (i.e., the presence of mediators, fluency, and explicit strategies).

Interestingly, learners did not accurately predict the mnemonic advantages of one-item categories compared to larger categories in Experiment 2. During both study and recall, category membership (and category size) may be much less salient for one-item categories than for larger categories. In one-item categories, learners may base mnemonic predictions on the particularities of individual categories, and the targets' prototypicality within that category, rather than on category size. Category size may only become a salient cue after learners see several items from that category. Further, in order to adequately adjust their predictions to accurately reflect their performance across the one-item categories, learners would need to average their recall across 12 different one-item categories. Keeping track of recall performance across 12 different categories (and averaging across those 12 items) may impose too great a cognitive load for learners to update their metacognitive predictions. Updating mnemonic predictions is often very difficult (Brigham & Pressley, 1988; Tullis & Benjamin, 2012), so it is impressive that learners can and do update their predictions to reflect the interaction between category size and expected test format for larger categories.

An intriguing aspect of our results is that category size is an extrinsic cue, and yet learners appropriately accounted for its effect in recall. Generally speaking, JOLs are much less sensitive to extrinsic cues than to intrinsic cues. For example, participants will provide a similar JOL for a 1-week retention interval as they will for a 1-day retention interval (Koriat, Bjork, Sheffer, & Bar, 2004) and believe that items presented once during study will be recalled just as well as items presented three times during study (Rhodes & Castel, 2008). Learners can account for extrinsic cues when making JOLs, but usually only when all experimental conditions are obvious to the individual (either by using a within-subject manipulation or explicitly describing all conditions; e.g., Castel, 2008; Koriat et al., 2004) and often requiring extensive practice or experience (but see Zechmeister & Shaughnessy, 1980). So why is category size such a powerful exception to this pattern of extrinsic cue neglect? One possibility is simply that humans are predisposed to detect covariation and relations in their environment (e.g., Allan, 1993). That is, relative to other extrinsic cues, relatedness is just particularly salient, such that learners are able to account for this factor while making immediate item-by-item JOLs. This explanation is merely speculative, and further research examining why categorical relations are appropriately accounted for in metacognitive judgments is needed.

Learners' metacognitive monitoring is very sophisticated. Not only do learners consider individual items' characteristics (e.g., Mueller et al., 2013), they also analytically consider multiple relations among studied items and the interaction between these relations and expected test format. Few studies have shown that expected test format affects the accuracy of metacognitive monitoring (Finley & Benjamin, 2012; Thiede, 1996). While much research has shown that relatedness plays a large role in multiple areas of cognition, the current studies show that learners' predictions about the influences of relatedness are incredibly accurate under various testing circumstances. These studies provide evidence of the complexity of learners' metacognitive monitoring: Not only do learners account for connections among stimuli when making predictions but they also further consider how these associations interact with test format.

Acknowledgments This research was in part supported by a Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery Grant to Kathleen L. Hourihan. Data were collected while Jonathan G. Tullis was a postdoctoral fellow supported by National Science Foundation REESE grant 0910218, and Institute of Education Sciences, U.S. Department of Education Grant # R305A1100060 awarded to R. Goldstone.

References

Allan, L. G. (1993). Human contingency judgments: Rule based or associative? *Psychological Bulletin, 114*(3), 435–448. doi:10.1037/0033-2909.114.3.435

Begg, I., Duft, S., Lalonde, P., Melnick, R., & Sanvito, J. (1989). Memory predictions are based on ease of processing. *Journal of Memory and Language, 28*(5), 610–632.

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

Benjamin, A. S. (2008). Memory is more than just remembering: Strategic control of encoding, accessing memory, and making decisions. *The Psychology of Learning and Motivation, 48*, 175–223. doi:IO.IOI6lSOO79.742J(07)48005-7

Benjamin, A. S., Bjork, R. A., & Schwartz, B. L. (1998). The mismeasure of memory: When retrieval fluency is misleading as a metamnemonic index. *Journal of Experimental Psychology: General, 127*(1), 55–68.

Benjamin, A. S., & Tullis, J. G. (2010). What makes distributed practice effective? *Cognitive Psychology, 61*(3), 228–247. doi:10.1016/j.cogpsych.2010.05.004

Borges, M. A., & Mandler, G. (1972). Effect of within-category spacing on free recall. *Journal of Experimental Psychology, 92*(2), 207–214.

Bousfield, W. A. (1953). The occurrence of clustering in the recall of randomly arranged associates. *Journal of General Psychology, 49,* 229.

Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision, 10,* 433–436. doi:10.1163/156856897X00357

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

Castel, A. D. (2008). Metacognition and learning about primacy and recency effects in free recall: The utilization of intrinsic and extrinsic cues when making judgments of learning. *Memory & Cognition, 36*(2), 429–437. doi:10.3758/mc.36.2.429

Dallett, K. M. (1964). Number of categories and category information in free recall. *Journal of Experimental Psychology, 68*(1), 1–12.

Einstein, G. O., & Hunt, R. R. (1980). Levels of processing and organization —Additive effects of individual-item and relational processing. *Journal of Experimental Psychology: Human Learning and Memory, 6*(5), 598. doi:10.1037/0278-7393.6.5.588

Finley, J. R., & Benjamin, A. S. (2012). Adaptive and qualitative changes in encoding strategy with experience: Evidence from the test-expectancy paradigm. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 38*(3), 632–652.

Gerjuoy, I. R., & Spitz, H. H. (1966). Associative clustering in free recall: Intellectual and developmental variables. *American Journal of Mental Deficiency, 70*(6), 918–927.

Hertzog, C., & Dunlosky, J. (2004). Aging, metacognition, and cognitive control. In B. H. Ross (Ed.), *The psychology of learning and motivation: Advances in research and theory* (Vol. 45, pp. 215–251). San Diego: Elsevier Academic Press.

Hertzog, C., Dunlosky, J., Robinson, A. E., & Kidder, D. P. (2003). Encoding fluency is a cue used for judgments about learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29*(1), 22–34. doi:10.1037/0278-7393.29.1.22

Hintzman, D. L. (2010). How does repetition affect memory? Evidence from judgments of recency. *Memory & Cognition, 38*(1), 102–115. doi:10.3758/MC.38.1.102

Hunt, R. R., & Einstein, G. O. (1981). Relational and item-specific information in memory. *Journal of Verbal Learning and Verbal Behavior, 20*(5), 497–514. doi:10.1016/s0022-5371(81)90138-9

Hunt, R. R., & Seta, C. E. (1984). Category size effects in recall: The roles of relational and individual item information. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 10*(3), 454–464.

Karpicke, J. D. (2009). Metacognitive control and strategy selection: Deciding to practice retrieval during learning. *Journal of Experimental Psychology: General, 138*(4), 469–486.

Kausler, D. H. (1974). *Psychology of verbal learning and memory.* New York: Academic Press.

Koriat, A. (1997). Monitoring one's own knowledge during study: A cue-utilization approach to judgments of learning. *Journal of Experimental Psychology: General, 126*(4), 349–370.

Koriat, A., Bjork, R. A., Sheffer, L., & Bar, S. K. (2004). Predicting one's own forgetting: The role of experience-based and theory-based processes. *Journal of Experimental Psychology: General, 133*(4), 643–656. doi:10.1037/0096-3445.133.4.643

Loftus, G. R., & Masson, M. E. J. (1994). Using confidence-intervals in within-subject designs. *Psychonomic Bulletin & Review, 1*(4), 476–490.

MacLeod, C. M., Pottruff, M. M., Forrin, N. D., & Masson, M. E. J. (2012). The next generation: The value of reminding. *Memory & Cognition, 40*(5), 693–702.

Matvey, G., Dunlosky, J., & Guttentag, R. (2001). Fluency of retrieval at study affects judgments of learning (jols): An analytic or nonanalytic basis for jols? *Memory & Cognition, 29*(2), 222–233.

Matvey, G., Dunlosky, J., & Schwartz, B. L. (2006). The effects of categorical relatedness on judgements of learning (jols). *Memory, 14*(2), 253–261. doi:10.1080/09658210500216844

Mueller, M. L., Tauber, S. K., & Dunlosky, J. (2013). Contributions of beliefs and processing fluency to the effect of relatedness on judgments of learning. *Psychonomic Bulletin & Review, 20*(2), 378–384. doi:10.3758/s13423-012-0343-6

Nelson, T. O., & Leonesio, R. J. (1988). Allocation of self-paced study time and the labor-in-vain effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14*(4), 676–686.

Nelson, T. O., & Narens, L. (1990). Metamemory: A theoretical framework and new findings. In G.H. Bower (Ed), *The psychology of learning and motivation* (Vol. 26, pp. 125–173). New York: Academic Press.

Nelson, T. O., & Narens, L. (1994). Why investigate metacognition? In J. Metcalfe & A. P. Shimamura (Eds.), *Metacognition: Knowing about knowing* (pp. 1–25). Cambridge: MIT Press.

Rhodes, M. G., & Castel, A. D. (2008). Memory predictions are influenced by perceptual information: Evidence for metacognitive illusions. *Journal of Experimental Psychology: General, 137*(4), 615–625. doi:10.1037/a0013684

Roenker, D. L., Thompson, C. P., & Brown, S. C. (1971). Comparison of measures for the estimation of clustering in free recall. *Psychological Bulletin, 76*(1), 45–48.

Ross, B. H., & Bradshaw, G. L. (1994). Encoding effects of remindings. *Memory & Cognition, 22,* 591–605.

Ross, B. H., & Kennedy, P. T. (1990). Generalizing from the use of earlier examples in problem solving. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16*(1), 42–55.

Ross, B. H., Perkins, S. J., & Tenpenny, P. L. (1990). Reminding-based category learning. *Cognitive Psychology, 22*(4), 460–492.

Rundus, D. (1971). Analysis of rehearsal processes in free recall. *Journal of Experimental Psychology, 89*(1), 63–77. doi:10.1037/H0031185

Slamecka, N. J. (1968). An examination of trace storage in free recall. *Journal of Experimental Psychology, 76*(4/1), 504–513. doi:10.1037/h0025695

Smith, S. M., Ward, T. B., Tindell, D. R., Sifonis, C. M., & Wilkenfeld, M. J. (2000). Category structure and created memories. *Memory & Cognition, 28*(3), 386–395. doi:10.3758/BF03198554

Tauber, S. K., & Dunlosky, J. (2012). Can older adults accurately judge their learning of emotional information? *Psychology and Aging, 27*(4), 924–933. doi:10.1037/a0028447

Thiede, K. W. (1996). The relative importance of anticipated test format and anticipated test difficulty on performance. *The Quarterly Journal of Experimental Psychology, Section A: Human Experimental Psychology 49*(4), 901–918. doi:10.1080/027249896392351

Thiede, K. W., Anderson, M. C. M., & Therriault, D. (2003). Accuracy of metacognitive monitoring affects learning of texts. *Journal of Educational Psychology, 95*(1), 66–73. doi:10.1037/0022-0663.95.1.66

Thiede, K. W., & Dunlosky, J. (1999). Toward a general model of self-regulated study: An analysis of selection of items for study and self-paced study time. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25*(4), 1024–1037.

Thiede, K. W., Wiley, J., & Griffin, T. D. (2011). Test expectancy affects metacomprehension accuracy. *British Journal of Educational Psychology, 81*(2), 264–273.

Tullis, J. G., & Benjamin, A. S. (2011). On the effectiveness of self-paced learning. *Journal of Memory and Language, 64*(2), 109–118.

Tullis, J. G., & Benjamin, A. S. (2012). The effectiveness of updating metacognitive knowledge in the elderly: Evidence from metamnemonic judgments of word frequency. *Psychology and Aging, 27,* 683–690.

Tullis, J. G., Finley, J. R., & Benjamin, A. S. (2013). Metacognition of the testing effect: Guiding learners to predict the benefits of retrieval. *Memory & Cognition, 41*, 492–442.

Tullis, J. G., Benjamin, A. S., & Liu, X. (2014a). Self-pacing study of faces of different races: Metacognitive control over study does not eliminate the cross-race recognition effect. *Memory & Cognition, 42*(6), 863–875.

Tullis, J. G., Benjamin, A. S., & Ross, B. H. (2014b). The reminding effect: Presentation of associates enhances memory for related

words in a list. *Journal of Experimental Psychology: General, 143*(4), 1526–1540.

Tullis, J. G., Braverman, M., Ross, B. H., & Benjamin, A. S. (2014c). Remindings influence the interpretation of ambiguous stimuli. *Psychonomic Bulletin & Review, 21*, 107–113.

Tullis, J. G., Benjamin, A. S., & Fiechter, J. (2015). *The efficacy of learners' testing choices*. Manuscript in preparation.

Tulving, E., & Pearlstone, Z. (1966). Availability versus accessibility of information in memory for words. *Journal of Verbal Learning and Verbal Behavior, 5,* 381–391.

Van Overschelde, J. P., Rawson, K. A., & Dunlosky, J. (2004). Category norms: An updated and expanded version of the battig and montague (1969) norms. *Journal of Memory and Language, 50*(3), 289–335. doi:10.1016/j.jml.2003.10.003

Wahlheim, C. N., Dunlosky, J., & Jacoby, L. L. (2011). Spacing enhances the learning of natural concepts: An investigation of mechanisms, metacognition, and aging. *Memory & Cognition, 39*(5), 750–763.

Wahlheim, C. N., & Jacoby, L. L. (2013). Remembering change: The critical role of recursive remindings in proactive effects of memory. *Memory & Cognition, 41*(1), 1–15.

Zechmeister, E. B., & Shaughnessy, J. J. (1980). When you know that you know and when you think that you know but you don't. *Bulletin of the Psychonomic Society, 5,* 41–44.