Published in final edited form as:

J Exp Psychol Learn Mem Cogn. 2017 June ; 43(6): 972–985. doi:10.1037/xlm0000360.

Test Expectancy and Memory for Important Information

Catherine D. Middlebrooks[1], **Kou Murayama**[2,3], and **Alan D. Castel**[1]

[1]University of California, Los Angeles

[2]University of Reading, UK

[3]Kochi University of Technology, Japan

Abstract

Prior research suggests that learners study and remember information differently depending upon the type of test they expect to later receive. The current experiments investigate how testing expectations impact the study of and memory for valuable information. Participants studied lists of words ranging in value from 1–10 points with the goal being to maximize their score on a later memory test. Half of the participants were told to expect a recognition test after each list, while the other half were told to expect a recall test. After several lists of receiving tests congruent with expectations, participants studying for a recognition test instead received an unexpected recall test. In Experiment 1, participants who had studied for a recognition test recalled less of the valuable information than participants anticipating the recall format. These participants continued to attend less to item value on future (expected) recall tests than participants who had only ever experienced recall testing. When the recognition tests were made more demanding in Experiment 2, value-based recall improved relative to Experiment 1: though memory for the valuable information remained superior when participants studied with the expectation of having to recall the information, there were no longer significant differences after accounting for recall testing experience. Thus, recall-based testing encouraged strategic, value-based encoding and enhanced retrieval of important information, while recognition testing in some cases limited value-based study and memory. These results extend prior work concerning the impact of testing expectations on memory, offering further insight into how people study important information.

Keywords

memory; value-directed remembering; test expectancy; test format; study skills/strategies

Whether information is successfully remembered at a later time is impacted by innumerable factors, but one critical component is the method by which such retrieval is tested. There are multiple ways in which to test one's memory—whether implicitly or explicitly—to determine the extent to which information has been encoded. Within the domain of explicit memory, testing is generally categorized as either recognition-based or recall-based. While there is debate as to the extent to which recall and recognition processes utilize similar mechanisms (e.g., Anderson & Bower, 1972; Carey & Lockhart, 1973; Mandler, 1980;

Please address all correspondence to Catherine D. Middlebrooks, Department of Psychology, University of California, Los Angeles, 1285 Franz Hall Box 951563, Los Angeles, CA 90095; (310) 206-9262; cmiddlebrooks@ucla.edu.

Rotello, Macmillan, & Van Tassel, 2000; Yonelinas & Parks, 2007; Wixted, 2007), the act of recognizing information seems to be qualitatively different than that of recalling it (Cabeza et al., 1997). These differences can have a notable impact on which information is later retrieved. Furthermore, the differences between recall and recognition, and the beliefs people have about these methods of retrieval, can also impact the encoding that takes place prior to any actual testing.

A number of studies indicate that participants who study with the expectation of an upcoming recall test outperform participants who study with the expectation of a recognition test on both recall and recognition tests (Balota & Neely, 1980; Hall, Grossman, & Elwood, 1976; Meyer, 1934, 1936; Neely & Balota, 1981; Schmidt, 1983; Thiede, 1996). In other words, studying with the expectation of an upcoming recall test can lead to better memory for the information and greater performance on the test, regardless of its actual format, than studying with the expectation of an upcoming recognition test.[1] These performance differences may partly stem from learners' beliefs about the demands of recognition-based test formats relative to recall-based formats. Indeed, learners generally expect recall tests to be more difficult than recognition tests (d'Ydewalle, Swerts, & De Corte, 1983; Hall et al., 1976; Murayama, 2005; Thiede, 1996) and expect to perform better on tests of recognition than recall (Speer & Flavell, 1979; Thiede, 1996). These expectations seem to be based not on experience with the specific tests in question, but rather with general experiences of recalling and recognizing information, as expectations of higher performance on recognition tests remain even when prior test performance suggests otherwise: Thiede (1996) reported that participants consistently provided higher judgments of learning (JOLs) when anticipating a more difficult recognition test than a less difficult recall test, despite having had prior experience with the tests and, thus, exposure to the difficulty.

These beliefs that learners hold about the differences in testing demands may encourage them to engage in entirely different encoding strategies as per the anticipated testing format. For instance, there is some evidence to suggest that anticipation of a recall test leads learners to engage in more associative encoding than anticipation of a recognition test (Anderson & Bower, 1972, 1974; Balota & Neely, 1980; Staresina & Davachi, 2006; but see Neely & Balota, 1981). This is consistent with educational reports: when anticipating a test in which they must simply recognize the correct answer, as in a multiple-choice exam, students emphasize detail-based memorization and a more unit-based focus; when anticipating a test in which they must produce the answer (e.g., an essay or short-answer exam), however, they endorse more holistic, associative strategies (e.g., drawing conceptual connections across multiple chapters) (e.g., Terry, 1933; but see Hakstian, 1971). Critically, such test-based differences in encoding seem to be intentional and strategic. For example, participants who studied a series of cue-target word pairs in anticipation of a free recall test (during which they need only recall the target words) reported intentionally ignoring the cue words, while

[1] Not all studies of test expectancy have consistently demonstrated this "recall superiority." A number of classroom-based studies and studies using more realistic study materials/tests (e.g., a multiple-choice exam based on text passages versus an old/new recognition test based on a list of unrelated words) have reported that expectancy-congruency between the anticipated test format and the received test is more important than the test format itself (for a review, see Lundberg & Fox, 1991). In laboratory contexts, however, recall superiority is the oft-reported finding. Additionally, both results (whether recall superiority or expectancy-congruent) are consistent with the notion that encoding itself is affected by expectations about the upcoming method of retrieval.

those anticipating a cued-recall test intentionally engaged in cue-target association strategies (Finley & Benjamin, 2012).

Altering one's encoding strategy to be more consistent with expectations of an upcoming test's demands should certainly not be considered ill-advised or inappropriate. On the contrary, "learning to the test," so to speak, suggests active metacognitive judgments and thoughtful self-regulation of one's study. The mistake on the part of the learner, however, would be in assuming that one's evaluations of the test demands are accurate or, perhaps worse, that high test performance owing to format-based strategizing during encoding is necessarily indicative of strong learning of the material. Implementing strategies during encoding specifically designed to match the anticipated demands of a given testing format may lead to high performance, consistent with a transfer-appropriate processing view (e.g., Morris, Bransford, & Franks, 1977), but high test performance does not necessarily mean that the learner has actually learned the material (Bjork, 1999; Funk & Dickson, 2011). The risk in assuming that performance reflects learning, or that one has accurately judged the demands of future retrieval situations, heightens when considering that to-be-remembered information in real-life situations often varies in terms of how important it is to remember. Oftentimes, it is not necessarily *how much* we remember, but *what* we remember that matters most, and any faulty expectations regarding testing demands, or misevaluations of true learning, could have more notable consequences if particularly important information ends up being forgotten.

When presented with enough to-be-remembered information that successfully recalling all of it is unlikely, research indicates that people can learn to be selective on the basis of value, attending specifically during study to the most important information at the expense of less important information (Castel, 2008; Castel, Benjamin, Craik, & Watkins, 2002; Castel, McGillivray, & Friedman, 2012; Middlebrooks, McGillivray, Murayama, & Castel, 2016; Middlebrooks, Murayama, & Castel, 2016). If the task demands are such that participants can (or at least *expect* to) remember most/all of the information, as when anticipating a recognition test (e.g., Shepard, 1967; Standing, Conezio, & Haber, 1970), selectively focusing on the subset of most important information and neglecting the less important information during study would seem not only unnecessary, but also rather counterproductive—why attempt to remember only a fraction of the information if you expect to remember it all?

The (anticipated) demands of recall tests, relative to recognition, may lead learners expecting to later receive a test of recall to adopt a value-directed, selective study strategy, whereas those expecting to receive a recognition-based test may forego such selectivity—learners studying for a recognition test should be markedly less likely to remember the most important information when given a surprise recall test than learners who studied with the expectation of having to later recall the information. Experiment 1 was designed to examine whether the aforementioned patterns of "recall superiority" in the test expectancy literature (e.g., Balota & Neely, 1980; Thiede, 1996) also extend to the study of and memory for valuable information specifically.

Experiment 1

Experiment 1 aimed to investigate the possible impact of test expectancy on encoding strategies that learners use when confronted with information that varies in value and the effects these strategies have on subsequent test performance. Prior research investigating the methods and mechanisms by which people study and remember valuable information have used tests of recall and recognition (e.g., Castel et al., 2002; Castel, Farb, & Craik, 2007; Middlebrooks, McGillivray, et al., 2016), but no work to date has investigated the impact that testing expectations themselves have on value-based learning.

An additional goal was to examine whether prior experience with one test format influences future value-based encoding and memory performance, even when the learner anticipates an alternative test format. Participants in Experiment 1 were explicitly told to expect either a recall test or a recognition test after each studied word list. After having studied four lists and receiving tests congruent with expectations, those participants who had been told to expect a recognition test instead received a surprise recall test following the fifth list. For subsequent lists, *all* participants were told to expect (and received) recall tests. Prior research indicates that test structure and content can guide strategy selection and use during future encoding (deWinstanley & Bjork, 2004; Garcia-Marques, Nunes, Marques, Carneiro, & Weinstein, 2015; Storm, Hickman, & Bjork, 2016), such that one learns what to study based on previous testing experience. In a similar vein, it may be the case that prior experience with a testing format that does not require a selective study strategy for optimum performance impacts subsequent strategy selection, despite changes in test format. It is presently unclear whether learners maintain prior format-based strategy use or whether they appropriately and successfully adapt to format changes.

Method

Participants—Participants consisted of 48 undergraduate students at the University of California, Los Angeles (38 female) ranging in age from 18 to 26 years (*M* = 20.4, *SD* = 2.1). Participants received partial credit for a course requirement.

Materials—The study was designed and presented to participants via the Collector program (Gikeymarcia/Collector, n. d.). Stimuli consisted of 8 lists containing 20 novel words apiece. Each of the words was randomly assigned a value ranging from 1 point to 10 points, with two words per list assigned to each value. The words in each list were randomly selected without replacement from a larger word bank of 280 random nouns (e.g., twig, button, point, brush). Word length ranged from 4–7 letters and averaged to 8.81 (*SD* = 1.57) on the log-transformed Hyperspace Analogue to Language (HAL) frequency scale[2], with a range from 5.48 to 12.65 (Lund & Burgess, 1996). The studied words were randomly selected from this bank for each participant in order to avoid any potential item effects (Middlebrooks, Murayama, et al., 2016; Murayama, Sakaki, Yan, & Smith, 2014). Thus, the words studied in List 1 for one participant might have been entirely different from another participant's

[2] The Log HAL frequency measure of the words included in the English Lexical Project ranges from 0 to 17, with an average frequency of 6.16 and a standard deviation of 2.40 (Balota et al., 2007).

List 1. Furthermore, one participant might study the word “twig” and another not, or might have studied “twig” as a 3-point word while another studied it as a 9-point word.

Procedure—Participants were told that they would be shown a series of word lists, each containing 20 different words. They were further told that each word would be paired with a value ranging from 1 to 10 points, that there would be two words per point value within each list, and that the words would be presented on the screen one at a time for 3 seconds apiece. Participants were instructed to remember as many of the words in each list as they could while also endeavoring to earn as many points as possible on a later test, with one’s score a sum of the points associated with each correctly remembered word.

Participants were randomly assigned to one of two testing conditions: an All Recall (All Rc) control group and a Recognition-then-Recall (RgRc) group. Participants in the All Rc group were told that they would be asked to recall the words from each list at the end of its presentation, at which point they would then be told their score (out of 110 possible points) and the number of words that they had successfully recalled. Thus, each of the 8 lists was followed by a recall test for participants in the All Rc group. Participants in the RgRc group were explicitly told to expect a recognition test at the conclusion of each list. They were told that each of the 20 studied words would be presented on the screen simultaneously along with 20 new words[3] (the words were arranged randomly on the screen) and that they would need to select the 20 items that they remembered having studied in the just-presented list. Importantly, participants *had* to select 20 items as having been studied before they could progress to the next list. In so doing, the proportion of items correctly recognized can be directly compared to the proportion of items correctly recalled (of the 20 items selected as “old,” how many were actually studied?) and false alarm rates are equivalent to miss rates. No words were reused across recognition tests (i.e., a word could not serve as a new word in multiple recognition tests and no studied words from prior lists were ever presented in later tests as new words). After selecting 20 items, participants were told their score (out of 110 possible points) and the number of words that they had successfully recognized.

Critically, participants in the RgRc group did not receive 8 recognition tests. Rather, they received recognition tests for Lists 1–4, but received recall tests for Lists 5–8. The recall test following List 5 was unexpected—participants studied List 5 with the expectation of receiving a recognition test based on prior instruction. After this surprise recall test, participants in the RgRc group were explicitly told to expect recall tests for the remainder of the task. So, while the recall test following List 5 was unexpected, the recall tests following Lists 6–8 were not (or were at least not intended to be) unexpected.

Results

Overall Memory Performance—The proportion of items correctly recalled or recognized, as applicable, across the 8 lists are provided in Table 1. Initial analyses were conducted to determine whether there were significant differences in overall recall,

[3] These “new” words were selected from the same 280-word word bank as the studied words. For any given participant in the RgRc group, words from the word bank had as much chance of being used as a studied word as of being used as a new word during a recognition test.

irrespective of item values, between the two testing conditions for Lists 5–8. A 2(Condition: All Recall or Recognition-then-Recall) × 4(List) repeated-measures ANOVA on total recall revealed a marginally significant List × Condition interaction, $F(3, 138) = 2.50$, $MSE = .01$, $\eta^2_G = .02$, $p = .06$. Participants in the RgRc group recalled a significantly smaller proportion ($M = .30$, $SD = .19$) of the items from List 5 than All Rc participants ($M = .42$, $SD = .16$), $t(46) = 2.38$, $d = 0.70$, $p = .02$. There were no significant condition differences in recall in Lists 6–8 (*d*s = 0.25, −0.06, and 0.26, respectively), *p*s > .38.

Thus, although participants' recall when expecting a recall test was superior to those expecting a recognition test (i.e., List 5), recall in the RgRc condition did not significantly differ from that of the All Rc condition in subsequent lists (i.e., Lists 6–8) (when all participants expected to receive a recall test), despite having had prior experience with a different testing format and having been exposed to additional items serving as "new" words during the recognition test. Consistent with prior research (e.g., Balota & Neely, 1980; Thiede, 1996), however, studying with the expectation of a recall test led to better memory for the material than studying with the expectation of a recognition test, despite recognition performance in Lists 1–4 having approached ceiling (see Table 1).

Value-Directed Remembering and Selectivity—To determine how value during study impacted subsequent recall, hierarchical linear modeling (HLM) was used to analyze recall as a function of the list in which an item appeared, its value, and the condition of study (Raudenbush & Bryk, 2002). HLM was used, as opposed to an ANOVA, for two primary reasons. Firstly, ANOVAs treat each value (or value bin, were values to be grouped) as a discrete entity rather than as part of a value continuum. ANOVAs cannot indicate whether there is a direct relationship between item value and recall and, thus, cannot thoroughly characterize any value-based strategies that may have been used during study. Without considering the value spectrum in its entirety, changes across lists in the impact of value on recall may be masked by such mean-based analytical techniques. Secondly, participants likely differ in *how* they strategically attend to value during study, so it would be inappropriate to analyze the data simply by comparing average recall across values, binned or otherwise. A participant who expects to remember many items, for example, may consider words worth 6 or more points worthy of attention during study, while a less confident participant may limit study to items worth only 9 or 10 points. In both cases, participants are executing a value-based strategy, limiting their study based on metacognitive judgments of personal memory capacity/ability. Simply comparing recall across value points without considering either the continuous structure of the value range or how participants may differ in their use of the values would prevent a complete understanding of value-directed remembering within the current task.

HLM accounts for within- and between-subject differences in strategy by first clustering the data within each participant, thereby accounting for individual differences in strategy, and *then* considering potential condition differences in the impact of value on recall, all while reflecting the to-be-remembered information as it was studied by participants and maintaining the overall data structure—a continuous value scale. (For more information on the use of HLM in investigating value-directed remembering, see Castel, Murayama, McGillivray, Friedman, & Link, 2013; Middlebrooks, McGillivray, et al., 2016; and Middlebrooks, Murayama, et al., 2016).

List 5: unexpected recall test: When studying List 5, participants in the RgRc condition expected to receive a recognition test as per pre-task instructions and prior experience in Lists 1–4. By presenting participants instead with a surprise recall test, it was possible to ascertain how (whether) participants attended to item value during study with the expectation of a recognition test format. Overall recognition performance to this point had been quite high, which could be interpreted as participants having had strong knowledge of the material. The question was whether participants had truly learned the high-value items or whether high recognition performance was actually masking low memory for the most important items. While participants may not have needed to be selective to perform well on the recognition tests (there is no cause to study selectively when one can perform well without having done so), did participants consider value at all when studying for a recognition test?

To compare value-based recollection between the two conditions, item-level recall performance in List 5 (based on a Bernoulli distribution, with 0 = *not recalled* and 1 = *recalled*; level 1 = items; level 2 = participants) was first modeled as a function of each item's value. Value was entered into the model as a group-mean centered variable, such that Value was anchored on the mean value point (5.50). The model further included the two study conditions as a level-2 predictor of those level-1 effects, with the Condition variable anchored on the All Rc control group (i.e., 0 = *All Recall*, 1 = *Recognition-then-Recall*). The model is essentially a logistic regression model with a dichotomous dependent variable, so the regression coefficients can be interpreted via their exponential (Raudenbush & Bryk, 2002). Specifically, exponential beta, Exp(B), is interpreted as the effect of the independent variable on the odds ratio of successful recall (i.e., the probability of recalling items divided by the probability of forgetting them) (Murayama et al., 2014). Exp(B) of more than 1.0 indicates a positive effect of the predictor, while an Exp(B) of less than 1.0 indicates a negative (or diminished) effect.

Recall performance in List 5 is presented in Figure 1a as a function of item value and condition. Table 2 reports the tested model and its estimated regression coefficients. There was a significant, positive effect of Value in the All Rc condition ($\beta_{10} = 0.31$, $p < .001$). In other words, participants in the All Rc condition were $e^{0.31} = 1.37$ times more likely to recall an item for one unit increases in its value. There was also a significant, negative interaction between Value and Condition ($\beta_{11} = -0.35$, $p < .001$), indicating that the effect of value was significantly weaker in the RgRc condition than the All Rc condition. Simple slope analysis revealed that, contrary to the All Rc condition, Value did not significantly predict recall probability in the RgRc condition ($\beta = -0.04$, $p = .35$)[4]. Thus, despite achieving high performance on the recognition tests, participants studying with the anticipation of a recognition-based test format did not learn the high-value information to the extent that

[4] The simple slope for the RgRc condition can be directly calculated by adding the β_{10} and β_{11} coefficients (i.e., $0.31+(-0.35) = -0.04$). To determine whether this slope is significant, the Condition predictor in the model was recoded, such that 0 = *Recognition-then-Recall* and 1 = *All Recall* (Hayes, 2013). Note that this was also done to determine the significance of any reported simple slopes hereafter.

might have been expected from their overall performance and seemed to have studied the items without attending to their general importance. While this inattention to value may have been a sufficient strategy for recognition, it was insufficient for true learning of the most important information, as evidence by the differences between conditions.

Lists 6–8: expected recall tests for everyone: The model used to investigate value-directed remembering in Lists 6–8, during which all participants studied with the expectation of a recall test, was similar to that of the model for List 5, the only differences being the inclusion of a List predictor (entered as a group-mean centered variable anchored on List 7) and a Value × List interaction predictor to level-1 of the model. Recall performance averaged across Lists 6–8 is presented in Figure 1b as a function of item value and condition. Table 3 reports the tested model and its estimated regression coefficients.

Value was a significantly positive predictor of recall performance in the All Rc condition ($\beta_{10} = 0.32$, $p < .001$). There was also a significant, negative interaction between Value and Condition ($\beta_{11} = -0.25$, $p < .001$), with the effect of value on recall probability being significantly weaker in the RgRc condition than in the All Rc condition. Simple slope analysis revealed that the effect of value in the RgRc condition, while significant and positive ($\beta = 0.08$, $p = .002$), was lower than in the All Rc condition. Thus, participants in the All Rc condition were $e^{0.32} = 1.38$ times more likely to recall an item for each one-unit increase in its value while participants in the RgRc condition were only $e^{0.08} = 1.08$ times more likely. The odds of All Rc participants recalling a 10-point item, for instance, were thus $e^{0.32* 10} = 24.53$ times greater than the odds of recalling a 1-point item, but only 2.23 times greater for RgRc participants.

There were no significant differences in recall as a function of List ($p = .21$), nor was there a List × Condition interaction ($p = .91$), as also indicated in the previously conducted repeated-measures ANOVA. There was not a significant interaction between List and Value in either the All Rc condition ($p = .11$) or the RgRc condition ($p = .19$), indicating that selectivity was constant across these final lists and condition differences in selectivity were also maintained.

Initial recall experience: Comparing value-directed remembering between conditions during Lists 6–8 reveals the differential effects that prior task experience had on participants' study strategies as a consequence of the manner in which their memory was to be tested. Namely, participants were less selective in their recall when their previous testing experience consisted of recognition tests than of recall tests. This would suggest that participants learned from testing and adjusted their strategies accordingly (Garcia-Marques et al., 2015; Jensen, McDaniel, Woodard, & Kummer, 2014; Storm et al., 2016). That RgRc participants were not as selective as the All Rc participants indicates that they did not learn to prioritize high-value items simply as a consequence of study design—such prioritization was unnecessary when their test performance was at ceiling—and thus did not have the same amount of practice executing a value-based strategy as those in the All Rc condition, for whom such a strategy was always important.

To determine whether the differences in selectivity between the two conditions was a consequence of the amount of practice in studying for a recall test (and thus the amount of practice utilizing a value-based study strategy), recall performance in Lists 6–8 by the RgRc condition was compared with recall performance in Lists 1–3 by the All Rc condition, depicted in Figure 1c. In so doing, performance in the first three lists of *expected* recall tests could be directly assessed. The same HLM model was used as when testing value-directed remembering in Lists 6–8, save that the list on which the List predictor was anchored was the second list for each condition (i.e., List 2 for the All Rc condition and List 7 for the RgRc condition). Table 3 reports the tested model and its estimated regression coefficients.

Value was a significantly positive predictor of recall performance in the All Rc condition ($\beta_{10} = 0.17$, $p < .001$) during Lists 1–3. Once again, there was a significant cross-level interaction between Value and Condition ($\beta_{11} = -0.09$, $p = .04$), such that the effect of value in the RgRc condition in Lists 6–8, while positive ($\beta\ = 0.08$, $p = .02$), was also significantly less than that of the All Rc control in Lists 1–3. There were no other significant effects (*p*s > .34). Thus, even after controlling for experience with studying for recall tests, participants who had previously studied for recognition tests were less attentive to the important items than participants without prior testing experience, recall or otherwise.

Discussion

The results of Experiment 1 are three-fold. Firstly, participants who studied in anticipation of a recognition test were far less attentive to item importance during study than those who studied in anticipation of a recall test. With average performance at ceiling, this lack of attention to value was not apparent from the recognition test performance but became quite evident in the surprise recall test following List 5. So, although it may have appeared that RgRc participants had learned those most important items during study, it was actually the case that participants who anticipated a recall test better learned the most important items than those anticipating a recognition test despite recalling fewer items overall than their counterparts could recognize.

Secondly, participants in the RgRc condition failed to recover from the shift in test formats. After it was made clear to RgRc participants that, while List 5 was surprising, Lists 6–8 would also now be followed by recall tests instead of recognition tests, they continued to be significantly less attentive to item importance than those in the All Rc condition. Notably, there was no evidence of changes in RgRc participants' attention to value across these three lists, indicating that their failure to attain selectivity comparable to that of All Rc participants was not only owing to a lack of practice—after all, the All Rc participants had already studied for and experienced multiple recall tests by this point in the task—but a failure to properly adapt to the demands of the recall format relative to the recognition format.

Thirdly, even after controlling for experience with studying for and taking recall tests, participants in the RgRc condition were *still* less selective than those in the All Rc condition. It would seem that having had earlier experience with recognition tests actually served to impair their selectivity during study, that prior experience with a test not requiring prioritization of high-value information impaired their ability to do so when later required.

Based on the results of Experiment 1, learning to be selective and strategic in the study of valuable information would seem to require not only experience with the material and general task (i.e., study 20 items varying in value and remember as many as possible while also maximizing one's score), but also experience with the testing format itself.

Experiment 2

In Experiment 1, prioritization of high-value information during study was evidently unnecessary for achieving high recognition test performance as participants were capable of correctly recognizing nearly all of the items without any special attendance to the most valuable. A selective study strategy, even in anticipation of a recognition test, however, may become a more sensible study choice were such high performance not so easily attained. Some research suggests that students modify their study based on the anticipated test demands (e.g., Dunlosky & Ariel, 2011; Entwistle & Entwistle, 2003; Garcia-Marques et al., 2015; Winne & Hadwin, 1998) rather than the testing format, per se. For instance, when anticipating a test which will necessitate deep processing, learners adopt deeper processing study strategies; likewise, anticipation of a test necessitating shallower processing leads to shallow processing during study (Ross, Green, Salisbury-Glennon, & Tollefson, 2006). Thus, test-expectancy effects on memory may be driven less by the test format itself than by judgments of the task demands associated with particular formats. The effect of anticipating one type of recognition test on study behaviors and memory performance might be quite different than another recognition test with different task demands, even when the to-be-remembered material itself is identical.

The recognition tests in Experiment 2 were made more difficult in order to investigate whether the evident effect of testing expectations on value-based learning demonstrated in Experiment 1 was driven by participants' general beliefs about testing formats or beliefs about the upcoming test's specific demands and difficulty, irrespective of format. Should expectations regarding the general format of the recognition test be more critical to encoding behaviors/choices than the demands of the *specific* recognition test with which they have been tasked, then value should continue to be a largely irrelevant factor to participants' study, and differences in List 5 selectivity/value-based recall should remain. On the other hand, participants in Experiment 2 should be more likely to recall high-value items than low-value items on the surprise recall test if value becomes a more relevant factor during study in light of a demanding recognition test.

In an attempt to create a more demanding recognition test condition, similarity was increased between the studied items and the unstudied lure items presented at test. Similarity was addressed in multiple ways, including semantic similarity (e.g., Elias & Perfetti, 1973; Underwood, 1965), pronunciation/acoustic similarity (e.g., Conrad, 1864; Kintsch & Buschke, 1969), and orthography (e.g., Logie, Della, Wynn, & Baddeley, 2000). The test items were also presented sequentially rather than simultaneously. Research concerning eyewitness memory suggests that simultaneous presentation during recognition testing can lead to more accurate identification of studied and novel information than sequential presentation because participants are better able to make comparisons between items when all possible selections are visible at the same time (e.g., Finley, Roediger, Hughes, Wahlheim, & Jacoby, 2015; Steblay, Dysart, & Wells, 2011).

Relatedly, research also suggests that forced-choice recognition formats in which participants can directly compare target and lure items, as in Experiment 1, rely more on mechanisms of familiarity than recollection (Holdstock et al., 2002; Kroll, Yonelinas, Dobbins, & Frederick, 2002), further increasing the retrieval demands at test. Recent work has demonstrated that correct recognition of previously studied valuable information is largely driven by a recollection component (Cohen, Rissman, Castel, Hovhannisyan, & Knowlton, 2015; Cohen, Rissman, Hovhannisyan, Castel, & Knowlton, under revision). Item importance has no apparent effect on recognition based on feelings of familiarity when the high-value information was not intentionally prioritized during study. In other words, there are no demonstrable differences in the recognition of low-and high-value information when learners can rely in feelings of familiarity in making their old/new judgments at test. When recognition is based on recollection, however, correctly recognized items tend to be the more valuable (Cohen et al., under revision). As such, participants may be less attentive to item value during study when the test demands are such that correct recognition can be easily achieved via feelings of familiarity, but more likely to adopt a value-based study strategy, *despite* the recognition-based format of the test, when correct recognition depends upon explicit recollection.

Method

Participants—Participants consisted of 48 undergraduate students at the University of California, Los Angeles (28 female) ranging in age from 18 to 26 years ($M = 20.4$, $SD = 1.8$). Participants received partial credit for a course requirement.

Materials—The materials used in Experiment 2 were very similar to those used in Experiment 1: there were 8 lists of 20 novel words with each word randomly assigned a value of 1–10 points and two words per list assigned to each value. The words in each list were randomly selected without replacement from a larger word bank of 200 random nouns. Word length ranged from 4–7 letters and averaged to 8.64 ($SD = 1.58$) on the log-transformed HAL frequency scale, with a range from 5.53 to 12.53.

Each of the words in the word bank used to select words for study was also paired with two lure words that were presented along with the corollary studied word during the recognition tests, as applicable. For instance, participants who studied the word “shovel” also saw the words “shove” and “hovel” during the recognition test; participants who studied the word “rain” saw the words “reign” and “train” at test. Lures were created in a number of different ways: some were homophones of the studied target item; others were homonyms, semantically similar words, words with similar spellings, etc. The lures were not created with the intention of systematically investigating whether error patterns differed depending on the alteration (e.g., are participants more likely to incorrectly select a target’s homonym than homophone?), but were simply intended to increase the test difficulty, such that anything but careful attendance to the items during study would leave a participant vulnerable to incorrectly selecting an item similar in sound, meaning, or appearance.

Procedure—The procedure used in Experiment 2 was identical to Experiment 1 save for the recognition tests. During the recognition tests, each of the 20 studied words was presented on the screen sequentially, rather than simultaneously, along with the 40 corollary new words (2 lures per studied word); participants were to select the 20 items that they remembered as having been just-presented during study. As in Experiment 1, participants were required to select 20 items as having been studied during the recognition test. Participants who had yet to select 20 items as having been studied before the test was finished were forced to select the *n* remaining items as "old." For example, if a participant had selected only 15 items as "old," and there were only 5 more items to be presented in the test, the option to indicate that an item was new was removed and participants were forced to select "old" for the 5 remaining items. Likewise, the option to indicate that items were "old" was removed in the event that a participant designated 20 items as "old" prior to the conclusion of the recognition test; participants were forced to select "new" for the remaining items. In addition to keeping hit rates comparable to recall accuracy, and false alarm rates equivalent to miss rates, requiring participants to select "old" or "new" in this manner ensured that participants who completed the recognition tests were consistently exposed to 60 items, keeping potential interference from new items presented during testing constant across RgRc participants.

Completion of the recognition tests was self-paced, but participants could not change their old/new response to an item once they had progressed to the next item. At the conclusion of the test, participants were told their score (out of 110 possible points) and the number of words that they had correctly recognized.

Results

Overall Recall Performance—The proportion of items correctly recalled or recognized, as applicable, across the 8 lists are provided in Table 1. As in Experiment 1, analyses were initially conducted to determine whether there were significant differences in overall recall, irrespective of item values, between the two testing conditions for Lists 5–8. A 2(Condition: All Recall or Recognition-then-Recall) × 4(List) repeated-measures ANOVA on total recall revealed a significant Condition × List interaction, $F(3, 138) = 3.01$, $MSE = .01$, $\eta^2_G = .02$, $p = .03$. While there were no significant changes in recall across lists for those in the All Rc condition ($p = .66$), there was a significant List effect in the RgRc condition, $F(3, 69) = 3.64$, $MSE = .01$, $\eta^2_G = .14$, $p = .02$, such that recall significantly increased from List 5 to List 8, *p*s < .033. Moreover, participants in the All Rc condition recalled significantly more items than the RgRc condition in Lists 5 and 6 (*d*s = 0.75 and 0.73, respectively; *p*s < .018), but there were no significant condition differences in recall for Lists 7 and 8 (*d*s = 0.09 and 0.25, respectively; *p*s > .39).

These results indicate that studying with the expectation of a recall test once again led to better memory for the to-be-remembered items than studying with the expectation of a recognition test, as demonstrated by the condition differences in List 5 recall. That the differences remained for List 6 recall may reflect effects of interference from the lengthier recognition tests relative to Experiment 1. Regardless, condition differences in overall recall were attenuated by List 7 (i.e., the RgRc group's third recall test), suggesting that any potential effect of interference was minimal or, at least, surmountable. RgRc participants' recall did not significantly differ from that of participants in the All Rc condition when expecting to receive a recall test in the final lists.

Value-Directed Remembering & Selectivity—All HLM analyses conducted to analyze the Experiment 2 data are identical to those used for Experiment 1.

List 5: unexpected recall test: Recall performance in List 5 is presented in Figure 2a as a function of item value and condition. Table 2 reports the tested model and its estimated regression coefficients. As in Experiment 1, there was a significant, positive effect of Value in the All Rc condition (β_{10} = 0.23, $p < .001$) and a significant cross-level interaction between Value and Condition (β_{11} = −0.14, p = .03)—participants were, once again, less selective in their study when anticipating a test of recognition than participants expecting a recall test. Unlike in Experiment 1, however, the effect of value in the RgRc condition was significantly positive (β = 0.09, p = .02). These results indicate that RgRc participants expecting to receive a recognition-based test in Experiment 2 *did* consider item importance during study, albeit notably less so than participants expecting to receive a recall test in the first place.

Lists 6–8: expected recall tests for everyone: Recall performance averaged across Lists 6–8 is presented in Figure 2b as a function of item value and condition. Table 3 reports the tested model and its estimated regression coefficients. As in Experiment 1, Value was a significantly positive predictor of recall probability in the All Rc condition (β_{10} = 0.19, $p < .001$). There was also a significant cross-level interaction between Value and Condition, (β_{11} = −0.11, p = .049), such that the positive effect of value in the RgRc condition (β = 0.08, p = .008) was significantly less than in the All Rc condition. So, while both groups were attentive to value in the final lists of the task, participants in the All Rc condition were more selective than those in the RgRc.

There were no significant differences in recall owing to List in the All Rc condition (p = .49), nor was there a List × Condition interaction (p = .16) for Lists 6–8. As in Experiment 1, there was also not a significant interaction between List and Value in the All Rc condition (p = .37), nor a three-way interaction between List, Value, and Condition (p = .31), indicating that condition differences were maintained across these final lists.

Initial recall experience: As in Experiment 1, performance in the first three lists of *expected* recall was directly compared between conditions (i.e., Lists 1–3 in the All Rc group versus Lists 6–8 in the RgRc group). Recall performance averaged across the first three anticipated recall tests is presented in Figure 2c as a function of item value and condition. Table 3 reports the tested model and its estimated regression coefficients. As in Experiment 1, Value was a significantly positive predictor of recall performance in the All Rc condition (β_{10} = 0.14, $p < .001$) during Lists 1–3. Contrary to Experiment 1, however, there was *not* a significant difference in the effect of value on recall between the All Rc condition and the RgRc condition (p = .14), indicating comparable selectivity during the first three lists of anticipated recall testing. There was a significant interaction between List and Value in the All Rc condition (β_{30} = 0.06, p = .045), such that selectivity improved across these first lists, consistent with prior work indicating the importance of task experience to the adoption of value-based study strategies (Castel, 2008; Castel et al., 2012; Middlebrooks, McGillivray, et al., 2016). This pattern was consistent in the RgRc condition, with no significant differences in the selectivity increase between conditions (*p* = .35).

Discussion

The demands of the recognition test format were increased in Experiment 2 relative to those of Experiment 1--instead of simply selecting the 20 studied items from a list of 40 items presented on the screen simultaneously, RgRc participants in Experiment 2 were to select the 20 studied items from a sequentially presented list of 60 items. Additionally, each of the studied items had two corollary lures in the recognition test, designed to be confusable with the studied item owing to similar pronunciations (e.g., racquet versus racket), spelling (e.g., stump versus stomp), meaning (e.g., bandage versus bandaid), and so forth.

Consistent with Experiment 1, participants expecting to receive a recognition test but who, in fact, received a recall test (List 5) were significantly less attentive to value during study than All Rc participants expecting to receive the recall test in the first place, despite having previously demonstrated strong recognition of the studied items, in general. Notably, however, item value had a positive effect on List 5 recall probability for RgRc participants in Experiment 2, whereas there was no such value effect on the surprise test for Experiment 1 RgRc participants. So, it would appear that the changes made to the recognition test in Experiment 2 were such that item value was now considered, or at least salient, during study, although to a lesser extent than in anticipation of a recall test.

Participants in the All Rc condition continued to study more selectively than those in the RgRc condition in Lists 6–8, during which both groups were told to expect (and received) free recall tests. These condition differences were negated, however, when experience with the recall test format itself was taken into account, directly contrasting the results of Experiment 1. So, contrary with Experiment 1 results, experience with the general task and study materials encouraged attendance to item value, though the extent of such selectivity depended upon the anticipated test format. Moreover, prior recognition testing did not hinder the appropriate adoption of a value-based study strategy in Experiment 2 as it did in Experiment 1.

That RgRc participants in Experiment 2 adapted to the change in test format from recognition to recall, and came to adopt value-based study strategies in anticipation of recall testing, suggests that the adjustments made to the recognition test in Experiment 2 relative to that which was administered during Experiment 1 encouraged changes in study and retrieval that enabled participants to more easily adapt to new formatting demands, as is further considered in the General Discussion.

General Discussion

When attempting to remember information, one of many contributing factors to successful memory at a later time is the method by which one expects the memories to be tested (e.g., Balota & Neely, 1980; Finley & Benjamin, 2012; Lundberg & Fox, 1991; Meyer, 1934, 1936; Murayama, 2006). Knowledge of the upcoming test format can affect both evaluations of encoding success (has this information been sufficiently learned?) (Thiede, 1996) and the particular behaviors/strategies in which learners engage during study (*how* should this information be learned?) (Finley & Benjamin, 2012; Garcia-Marques et al., 2015; Terry, 1933, 1934). The current experiments examined how expectations regarding the upcoming test format can affect one's study of information varying in value or importance—are learners similarly likely to learn and remember important information when anticipating a recall test as when anticipating a test of recognition? In light of differences in value-based study between recall and recognition test formats in Experiment 1, Experiment 2 aimed to clarify how encoding and during study may differ owing to the test's specific demands, rather than the general format, and the extent to which the likelihood of adopting and successfully executing strategies appropriate for one test format is influenced by prior experience with an alternate format.

In both experiments, participants who studied in anticipation of a recognition test were far less likely to remember the most important items in a surprise recall test than those participants who studied with the expectation of a recall test, consistent with research indicating better memory overall when studying with the expectation of a recall-based than recognition-based test (e.g., Balota & Neely, 1980; Thiede, 1996). Notably, this pattern of "recall superiority" was evident despite the fact that prior recognition test performance was quite high and would have otherwise suggested sufficient knowledge of the important information. Thus, it would seem that high recognition performance in the current experiments largely masked poorer learning of the valuable material relative to participants expecting to receive a recall test. Had only recognition tests been administered, it would have appeared as though the most important information had been effectively learned when, in fact, memory was significantly inferior to that of participants studying for a recall test (Funk & Dickson, 2011).

Additionally, participants in both experiments who had previously received recognition tests, but were told to now expect recall tests for the remainder of the task, were significantly less selective in their study of the valuable information than participants who had only ever studied for and received recall tests, indicating the importance of experience with not only the material and general task structure, but also the method of testing itself. The way in which participants chose to study was not purely dependent upon the materials themselves, but rather on how they would later be asked to retrieve the information—participants learned what and how to learn based on the test format (cf., Finley & Benjamin, 2012; Garcia-Marques et al., 2015; Jensen et al., 2014).

It is possible that RgRc participants failed to adjust their selectivity in light of the recall tests because they doubted the veracity of the experimenter's instruction to expect recall tests for the remainder of the task, given that prior expectations of recognition testing were a consequence of similarly explicit instruction provided at the start of the task and yet had been violated in List 5. Participants may have been hesitant to accommodate this change in instruction, believing it possible that the format would unexpectedly switch back to recognition again (or to an entirely novel format). There is, however, reason to suspect that a lack of trust in experimenter instruction does not explain the present results. Firstly, participants could clearly recognize more items than they could recall. Had participants seriously entertained the possibility that a recognition test might, at some point, be administered during Lists 6–8, despite instructions to the contrary, they still should have prepared for recall—being able to recognize an item does not guarantee accurate recall (which would have become evident during the List 5 recall test), but one can surely recognize an item which can also be recalled.

A perhaps more convincing argument against distrust motivating selectivity differences, though, is based on the differences seen between Experiment 1 and 2 with respect to RgRc participants' adaptation to the recall tests. Experiment 1 participants with prior recognition testing experience were significantly less selective than participants with experience only of recall testing; despite having studied six lists of item-value pairings, RgRc participants were still less attentive to important information and less able to remember it than All Rc participants with entirely no task experience. This was not the case in Experiment 2. Although less selective in Lists 6–8 than All Rc participants, Experiment 2 participants were not less selective once accounting for prior experience with recall testing specifically. If a lack of trust motivated the persistent differences in selectivity between the RgRc and All Rc participants in Experiment 1, it should have similarly done so in Experiment 2. There is no reason to believe that Experiment 1 participants were so much less trusting than participants in Experiment 2, to the point that they failed to adapt to the recall test format.

The differential impact of the recognition test format relative to the recall test format on selectivity between Experiments 1 and 2 is, however, consistent with the notion that testing expectation effects on strategy adoption and encoding behavior are less a consequence of the expectations learners have about what recognition- and recall-based formats *generally* entail, but their expectations regarding the inherent demands of the specific test which they are to receive. Learners seem to hold broad beliefs about the demands and relative difficulty levels of recognition-based and recall-based tests (Terry, 1933), but the results of the current experiments suggest that they can modify their study based on continued experience with the specific demands of the test.

In the absence of any other indicators regarding the demands of an upcoming recognition test, learners may generally believe that a feeling of familiarity at test will be an efficacious determinant of correct recognition—along the lines of "I'll know it when I see it" (Terry, 1933). Widely endorsed dual-process models, however, clearly outline both recollection and familiarity components of recognition memory (cf. Yonelinas, 2002). Whereas familiarity alone might have been sufficient for correctly recognizing "plane" and rejecting "drizzle" in Experiment 1, it was likely insufficient for the correct recognition of "plane" and correct rejection of "plain" in Experiment 2 (Gallo, 2004; Holdstock et al., 2002; Schmid, Herholz, Brandt, & Buchner, 2010). This is not to say that recognition in Experiment 1 would never have been based on recollection, or that feelings of familiarity never contributed to recognition in Experiment 2. After all, had RgRc participants in Experiment 2 relied purely on recall mechanisms while completing the recognition tests, their overall recall performance on the unexpected List 5 recall test would likely not have been significantly lower than that of All Rc participants. The demands of the recognition test in Experiment 2 were, nonetheless, such that correct recognition very likely depended more heavily on explicit recollection than in Experiment 1 (Holdstock et al., 2002; Kroll et al., 2002).

The demonstrated attention to value on the (surprise) List 5 recall test by RgRc participants in Experiment 2, but not Experiment 1, is further consistent with recent work indicating that recognition driven by explicit recollection is more likely to be value-based than recognition driven by feelings familiarity, even in the absence of intentional strategizing (Cohen et al., under revision). Extending this finding, the absence of a value effect in Experiment 1 by RgRc participants indicates not only a lack of value-based study strategizing when anticipating recognition testing, but also implies that prior recognition performance in Lists 1–4 was not (primarily) driven by recollection—recognition based on explicit recollections of the studied items would have been enhanced by the value of the items themselves and, thus, resulted in some degree of a value effect on the surprise recall. The List 5 value effect exhibited by RgRc participants in Experiment 2, however, is consistent with the supposition that recognition was more greatly aided by recollection than in Experiment 1. Importantly, RgRc participants were only twice as likely to recall a 10-point word as a 1-point word, whereas All Rc participants were approximately 25 times as likely. So while this effect of value on recall was significant for RgRc participants, the small magnitude suggests more automatic effects of value on recollection-based recognition memory than value-based strategizing during encoding (Cohen et al., under revision).

The differences between RgRc performance relative to All Rc performance in Experiments 1 and 2 cannot, however, be solely explained by the possible differences in the dominant mechanism (whether familiarity or recollection) underlying their recognition performance. This may account for the differences in List 5 recall, but it does not completely elucidate why differences between the RgRc and All Rc conditions perpetuated, even after accounting for recall experience, in Experiment 1 but not Experiment 2. It may be that the overall design of the recognition test in Experiment 2 made participants more keenly aware of ways in which they could potentially misremember items or confuse the studied and unstudied items at test. This knowledge may have led RgRc participants in Experiment 2 to engage in deeper or more elaborative encoding strategies (Craik & Lockhart, 1972; Craik & Tulving, 1975) in order to emphasize defining characteristics of the items and later aid in differentiating the studied from unstudied (which may also have made the items more distinctive as a consequence; Gallo, Meadow, Johnson, & Foster, 2008).[5]

Recent work concerning value-based learning and selectivity indicates that the study of high-value information relative to low-value information, in anticipation of a recall test, is associated with greater activity in regions of semantic processing (Cohen, Rissman, Suthana, Castel, & Knowlton, 2014, 2016). If RgRc participants' recognition in Experiment 2 was based more on recollection than familiarity judgments, this could have made the transition to a purely recall-based test format less jarring, in that participants could have adapted encoding strategies already being used in anticipation of recognition testing for the recall tests. Although RgRc participants in Experiment 2 may not have intentionally encoded high-

[5] There have been numerous studies to suggest that such deep encoding can actually encourage false recall and recognition of critical lures (e.g., Dodd & MacLeod, 2004; Rhodes & Anastasi, 2000; Thapar & McDermott, 2001; Toglia, Neuschatz, & Goodwin, 1999). Importantly, these studies have predominantly relied on the DRM paradigm, in which there is a single critical lure (e.g., "sleep") associated with a given set of studied items ("bed," "rest", "dream," etc.) (Deese, 1959; Roediger & McDermott, 1995), which differs from the current set of experiments. In Experiment 1, participants were required to distinguish studied items from unrelated items; in Experiment 2, studied items were associated with two lures during test, but the lures were related only to a single studied item and were not lures as a consequence of the studied list in its entirety, or even a substantial subset of the studied list.

value items specifically to a deeper extent than low-value items when expecting recognition tests (Cohen et al., under revision), it would have been conceivably easier to incorporate additional aspects of the material, like value, in their study if they were already utilizing deeper encoding strategies/processing. RgRc participants in Experiment 1 may not have studied in a manner that could be optimally adapted to recall-based testing, hence the struggle to study with the goal of being able to produce the items at test *and* remember the most valuable ones.

Although there is good reason to suspect that the adjustments made to the recognition test in Experiment 2 resulted in differences with respect to how recognition was realized, whether via recollection or familiarity (Holdstock et al., 2002; Kroll et al., 2002), it cannot be confirmed based on the design of the current experiments. Future research could test this notion using more direct tests of recollection and familiarity (e.g., remember/know judgments; Tulving, 1985). Robust selectivity in spite of prior recognition testing experience, as in Experiment 2, would indicate that recognition tests which necessitate recollection or explicit retrieval of studied items could be less damaging to the future adoption of appropriate study strategies when full retrieval, as in a free recall test, is necessary.

Future research should also investigate the extent to which general test difficulty contributed to condition differences. Although the recognition test in Experiment 2 was designed to be more demanding than that of Experiment 1, the recall test format was arguably the most demanding. Rather than the selectivity differences being driven by the extent to which explicit recollection was required at test, they may instead have been driven by the more general differences in task demand. Including a modified All Rc condition in which Lists 1–4 are followed by an easy recall test relative to Lists 5–8 might help to qualify the impact that the retrieval mechanisms and broader cognitive demands of the task had on selectivity and value-based study. A failure to adapt to more challenging recall tests after experiencing easy recall tests would suggest that the differences in the current experiments arose from general demands, rather than format-based demands or characteristics. Alternatively, swift adaptation from an easy to difficult recall tests would highlight not only the importance of common retrieval mechanisms in adapting to changing test demands, but also the importance of anticipating recollection-based testing formats when studying valuable information.

The decision to engage in a selective, value-based strategy, to prioritize high-value items over less valuable items when all cannot be remembered, reflects an active monitoring of one’s capacity limitations and the effectiveness of alternative study attempts and strategies. As such, it may be that asking learners who are anticipating tests which are less likely to stimulate selective study, such as the recognition tests in the current experiments, to make predictions regarding how well they will remember the important information, or to provide JOLs during study, would help to overcome any detrimental effects that testing expectations might have on the consideration of item importance. Doing so could make the effectiveness (or lack thereof) of any non-value-based study strategies more salient to learners, particularly with continued task experience (Hertzog, Price, & Dunlosky, 2008), thus encouraging value-based study in spite of test format.

Given the current findings, future research should also investigate whether expectations of either recognition-based or recall-based testing alter learners' attention to pedagogical importance of more realistic study materials (e.g., text passages) and testing formats, such as multiple-choice or essay exams (Lundberg & Fox, 1991; McDaniel, Blischak, & Challis, 1994; Murayama, 2003). Students may be less likely to consider material importance when preparing for tests of recognition, as suggested by the current experiments, but more likely to do so when preparing for open-response, recall-based exams (Rickards & Friedman, 1978). In addition, consideration should also be given to whether testing expectations influence how learners allocate study time as a function of item importance (cf., Ariel, Dunlosky, & Bailey, 2009; Middlebrooks, Murayama, et al., 2016). While it cannot be said that the influence recognition tests had in the current experiments will be the same for multiple-choice tests or other recognition-based formats, the results do suggest that knowledge of the upcoming format and its design, as well as prior experience with alternate formats, can impact the encoding of important information.

Summary—Prior research has demonstrated that expectations about upcoming tests affect metacognitive evaluations of learning, strategy selection and execution, and general memory performance. The current experiments extend these findings to address how testing expectations influence the study of and memory for important information. The results of the present research indicate that, in some situations, experience with recognition testing can prove injurious to value-based selectivity during study and subsequent memory for the most important information as compared with free recall testing. This negative influence of recognition testing was, however, mitigated by altering the demands of the recognition test, with some indication that the anticipation of a testing format that relies more heavily on explicit recollection of the to-be-remembered material is more conducive to encouraging selective attendance to important information during study and, thus, better memory at test than formats in which explicit recollection is less necessary.

Acknowledgments

The authors thank Elizabeth Bjork and Robert Bjork for helpful comments; Tyson Kerr for help with the experimental design and data collection; and Brenna McCarty and Cassia Ng for help with data collection. This work was supported in part by the National Institutes of Health (National Institute on Aging), Award Number R01AG044335 (to A. Castel). This work was further supported by the Marie Curie Career Integration Grant (CIG630680; to K. Murayama) and JSPS KAKENHI (Grant number 15H05401; to K. Murayama).

References

Anderson JR, Bower GH. Recognition and retrieval processes in free recall. Psychological Review. 1972; 79:97–123.

Anderson JR, Bower GH. A propositional theory of recognition memory. Memory & Cognition. 1974; 2:406–412. [PubMed: 21274765]

Ariel R, Dunlosky J, Bailey H. Agenda-based regulation of study-time allocation: When agendas override item-based monitoring. Journal of Experimental Psychology: General. 2009; 138(3):432–447. [PubMed: 19653800]

Balota DA, Yap MJ, Cortese MJ, Hutchison KA, Kessler B, Loftis B, Neely JH, Nelson DL, Simpson GB, Treiman R. The English Lexicon Project. Behavior Research Methods. 2007; 39:445–459. [PubMed: 17958156]

Balota DA, Neely JH. Test-expectancy and word-frequency effects in recall and recognition. Journal of Experimental Psychology: Human Learning and Memory. 1980; 6:576–587.

Bjork, RA. Assessing your own competence: Heuristics and illusions. In: Gopher, D., Koriat, A., editors. Attention and performance XVII: Cognitive regulation of performance: Interaction of theory and application. Cambridge, MA: MIT Press; 1999. p. 435-459.

Cabeza R, Kapur S, Craik FIM, McIntosh AR, Houle S, Tulving E. Functional neuroanatomy of recall and recognition: A PET study of episodic memory. Journal of Cognitive Neuroscience. 1997; 9:254–265. [PubMed: 23962015]

Carey ST, Lockhart RS. Encoding differences in recognition and recall. Memory & Cognition. 1973; 1:297–300. [PubMed: 24214561]

Castel AD. The adaptive and strategic use of memory by older adults: Evaluative processing and value-directed remembering. The Psychology of Learning and Motivation. 2008; 48:225–270.

Castel AD, Benjamin AS, Craik FIM, Watkins MJ. The effects of aging on selectivity and control in short-term recall. Memory & Cognition. 2002; 30:1078–1085. [PubMed: 12507372]

Castel AD, Farb NAS, Craik FIM. Memory for general and specific information in younger and older adults: Measuring the limits of strategic control. Memory & Cognition. 2007; 35:689–700. [PubMed: 17848027]

Castel, AD., McGillivray, S., Friedman, MC. Metamemory and memory efficiency in older adults: Learning about the benefits of priority processing and value-directed remembering. In: Naveh-Benjamin, M., Ohta, N., editors. Memory and aging: Current issues and future directions. New York: Psychology Press; 2012. p. 245-270.

Castel AD, Murayama K, Friedman MC, McGillivray S, Link I. Selecting valuable information to remember: Age-related differences and similarities in self- regulated learning. Psychology & Aging. 2013; 28:232–242. [PubMed: 23276210]

Cohen, MS., Rissman, J., Castel, AD., Hovhannisyan, M., Knowlton, BJ. Dual process dissociations reveal how free recall tests interspersed during learning can bolster the efficacy of value- driven strategy use. Poster presented at the 56th annual meeting of the Psychonomic Society; Chicago, IL. 2015.

Cohen MS, Rissman J, Hovhannisyan M, Castel AD, Knowlton BJ. Dissociating strategy-driven and automatic effects of value on memory. (under revision).

Cohen MC, Rissman J, Suthana NA, Castel AD, Knowlton BJ. Value-based modulation of memory encoding involves strategic engagement of front-temporal semantic processing regions. Cognitive, Affective, & Behavioral Neuroscience. 2014; 14:578–592.

Cohen MC, Rissman J, Suthana NA, Castel AD, Knowlton BJ. Effects of aging on value-directed modulation of semantic network activity during verbal learning. NeuroImage. 2016; 125:1046–1062. [PubMed: 26244278]

Conrad R. Acoustic confusions in immediate memory. British Journal of Psychology. 1964; 55:75–84. doi: j.2044-8295.1964.tb00899.x.

Craik FIM, Lockhart RS. Levels of processing: A framework for memory research. Journal of Verbal Learning and Verbal Behavior. 1972; 11:671–684.

Craik FIM, Tulving E. Depth of processing and the retention of words in episodic memory. Journal of Experimental Psychology: General. 1975; 104:263–294.

Deese J. On the prediction of occurrence of particular verbal intrusions in immediate recall. Journal of Experimental Psychology. 1959; 58:17–22. [PubMed: 13664879]

deWinstanley PA, Bjork EL. Processing strategies and the generation effect: Implications for making a better reader. Memory & Cognition. 2004; 32:945–955. [PubMed: 15673182]

Dodd MD, MacLeod CM. False recognition without intentional learning. Psychonomic Bulletin & Review. 2004; 11:137–142. [PubMed: 15116999]

Dunlosky J, Ariel R. Self-regulated learning and the allocation of study time. Psychology of Learning & Motivation. 2011; 54:103–140.

d'Ydewalle G, Swerts A, De Corte E. Study time and test performance as a function of test expectations. Contemporary Educational Psychology. 1983; 8:55–67.

Elias CS, Perfetti CA. Encoding task and recognition memory: The importance of semantic encoding. Journal of Experimental Psychology. 1973; 99:151–156.

Entwistle N, Entwistle D. Preparing for examinations: The interplay of memorizing and understanding, and the development of knowledge objects. Higher Education Research and Development. 2003; 22:19–41.

Finley JR, Benjamin AS. Adaptive and qualitative changes in encoding strategy with experience: Evidence from the test-expectancy paradigm. Journal of Experimental Psychology: Learning, Memory, & Cognition. 2012; 38:632–652.

Finley JR, Roediger HL, Hughes AD, Wahlheim CN, Jacoby LJ. Simultaneous versus sequential presentation in testing recognition memory for faces. American Journal of Psychology. 2015; 128:173–195. [PubMed: 26255438]

Funk SC, Dickson KL. Multiple-choice and short-answer exam performance in a college classroom. Teaching of Psychology. 2011; 38:273–277.

Gallo DA. Using recall to reduce false recognition: Diagnostic and disqualifying monitoring. Journal of Experimental Psychology: Learning, Memory, & Cognition. 2004; 30:120–128.

Gallo DA, Meadow NG, Johnson EL, Foster KT. Deep levels of processing elicit a distinctiveness heuristic: Evidence from the criterial recollection task. Journal of Memory and Language. 2008; 58:1095–1111.

Garcia-Marques L, Nunes LD, Marques P, Carneiro P, Weinstein Y. Adapting to test structure: Letting testing teach what to learn. Memory. 2015; 23:365–380. [PubMed: 24568583]

Gikeymarcia/Collector. [Retrieved March 13, 2016] (n. d.) from https://github.com/gikeymarcia/Collector.

Hakstian RA. The effects of type of examination anticipated on test preparation and performance. The Journal of Educational Research. 1971; 64:319–324.

Hall JW, Grossman LR, Elwood KD. Differences in encoding for free recall vs. recognition. Memory & Cognition. 1976; 4:507–513. [PubMed: 21286974]

Hayes, AF. Mediation, Moderation, and Conditional Process Analysis. New York: Guilford Press; 2013.

Hertzog C, Price J, Dunlosky J. How is knowledge generated about memory encoding strategy effectiveness? Learning and Individual Differences. 2008; 18:430–445. [PubMed: 19043596]

Holdstock JS, Mayes AR, Roberts N, Cezayirli E, Isaac CL, O'Reilly RC, Norman KA. Under what conditions is recognition spared relative to recall after selective hippocampal damage in humans? Hippocampus. 2002; 12:341–351. [PubMed: 12099485]

Jensen JL, McDaniel MA, Woodard SM, Kummer TA. Teaching to the test…or testing to teach: Exams requiring higher order thinking skills encourage greater conceptual understanding. Educational Psychology Review. 2014; 26:307–329.

Kintsch W, Buschke H. Homophones and synonyms in short-term memory. Journal of Experimental Psychology. 1969; 80:403–407.

Kroll NEA, Yonelinas AP, Dobbins IG, Frederick CM. Separating sensitivity from response bias: Implications of comparisons of yes-no and forced-choice tests for models and measures of recognition memory. Journal of Experimental Psychology: General. 2002; 131:241–254. [PubMed: 12049242]

Logie RH, Della SS, Wynn V, Baddeley AD. Visual similarity effects in immediate verbal serial recall. Quarterly Journal of Experimental Psychology Section A. 2000; 53:626–646.

Lund K, Burgess C. Producing high-dimensional semantic spaces from lexical co-occurrences. Behavior Research Methods, Instruments, & Computers. 1996; 28:203–208.

Lundberg MA, Fox PW. Do laboratory findings on test expectancy generalize to classroom outcomes? Review of Educational Research. 1991; 61:94–106.

Mandler G. Recognizing: The judgment of previous occurrence. Psychological Review. 1980; 87:252–271.

McDaniel MA, Blischak DM, Challis B. The effects of test expectancy on processing and memory of prose. Contemporary Educational Psychology. 1994; 19:230–248.

Meyer G. An experimental study of the old and new types of examination: The effect of the examination set on memory. Journal of Educational Psychology. 1934; 25:641–661.

Meyer G. The effect of recall and recognition on the examination set in classroom situations. Journal of Educational Psychology. 1936; 27:81–99.

Middlebrooks CD, McGillivray S, Murayama K, Castel AD. Memory for allergies and health foods: How younger and older adults strategically remember critical health information. Journals of Gerontology, Series B: Psychological Sciences and Social Sciences. 2016; 71:389–399.

Middlebrooks CD, Murayama K, Castel AD. The value in rushing: Memory and selectivity when short on time. Acta Psychologica. 2016; 170:1–9. [PubMed: 27305652]

Morris CD, Bransford JD, Franks JJ. Levels of processing versus transfer appropriate processing. Journal of Verbal Learning and Verbal Behavior. 1977; 16:519–533.

Murayama K. Test format and learning strategy use. Japanese Journal of Educational Psychology. 2003; 51:1–12.

Murayama K. Exploring the mechanism of test-expectancy effects on strategy change. Japanese Journal of Educational Psychology. 2005; 53:172–184.

Murayama K. “Adaptation to the test”: A review of problems and perspectives. Japanese Journal of Educational Psychology. 2006; 54:265–279.

Murayama K, Sakaki M, Yan VX, Smith G. Type-1 error inflation in the traditional by-participant analysis to metamemory accuracy: A generalized mixed-effects model perspective. Journal of Experimental Psychology: Learning, Memory, & Cognition. 2014; 40:1287–1306.

Neely JH, Balota DA. Test-expectancy and semantic-organization effects in recall and recognition. Memory & Cognition. 1981; 9:283–300.

Raudenbush, SW., Bryk, AS. Hierarchical linear models: Applications and data analysis methods. second. Newbury Park, CA: Sage; 2002.

Rickards JP, Friedman F. The encoding versus the external storage hypothesis in note taking. Contemporary Educational Psychology. 1978; 3:136–143.

Rhodes MG, Anastasi JS. The effects of a levels-of-processing manipulation on false recall. Psychonomic Bulletin & Review. 2000; 7:158–162. [PubMed: 10780030]

Roediger HL III, McDermott KB. Creating false memories: Remembering words not presented in lists. Journal of Experimental Psychology: Learning, Memory, & Cognition. 1995; 21:803–814.

Ross ME, Green SB, Salisbury-Glennon JD, Tollefson N. College students’ study strategies as a function of testing: An investigation into metacognitive self-regulation. Innovative Higher Education. 2006; 30:361–375.

Rotello CM, Macmillan NA, Van Tassel G. Recall-to-reject in recognition: Evidence from ROC curves. Journal of Memory and Language. 2000; 43:67–88.

Schmid J, Herholz SC, Brandt M, Buchner A. Recall-to-reject: The effect of category cues on false recognition. Memory. 2010; 18:863–882. [PubMed: 21108106]

Schmidt SR. The effects of recall and recognition test expectancies on the retention of prose. Memory & Cognition. 1983; 11:172–180. [PubMed: 6865751]

Shepard R. Recognition memory for words, sentences, and pictures. Journal of Verbal Learning and Verbal Behavior. 1967; 6:156–163.

Speer JR, Flavell JH. Young children’s knowledge of the relative difficulty of recognition and recall memory tasks. Developmental Psychology. 1979; 15:214–217.

Standing L, Conezio J, Haber RN. Perception and memory for pictures: Single-trial learning of 2500 visual stimuli. Psychonomic Science. 1970; 19:73–74.

Staresina BP, Davachi L. Differential encoding mechanisms for subsequent associative recognition and free recall. The Journal of Neuroscience. 2006; 26:9162–9172. [PubMed: 16957073]

Steblay NK, Dysart JE, Wells GL. Seventy-two tests of the sequential lineup superiority effect: A meta-analysis and policy discussion. Psychology, Public Policy, and Law. 2011; 17:99–139.

Storm BC, Hickman ML, Bjork EL. Improving encoding strategies as a function of test knowledge and experience. Memory & Cognition. 2016; 44:660–670. [PubMed: 26822535]

Terry PW. How students review for objective and essay tests. The Elementary School Journal. 1933; 33:592–603.

Terry PW. How students study for three types of objective tests. The Journal of Educational Research. 1934; 27:333–343.

Thapar A, McDermott KB. False recall and false recognition induced by presentation of associated words: Effects of retention interval and level of processing. Memory & Cognition. 2001; 29:424–432. [PubMed: 11407419]

Thiede KW. The relative importance of anticipated test format and anticipated test difficulty on performance. The Quarterly Journal of Experimental Psychology. 1996; 49A:901–918.

Toglia MP, Neuschatz JS, Goodwin KA. Recall accuracy and illusory memories: When more is less. Memory. 1999; 7:233–256. [PubMed: 10645381]

Tulving E. Memory and consciousness. Canadian Psychology. 1985; 26:1–12.

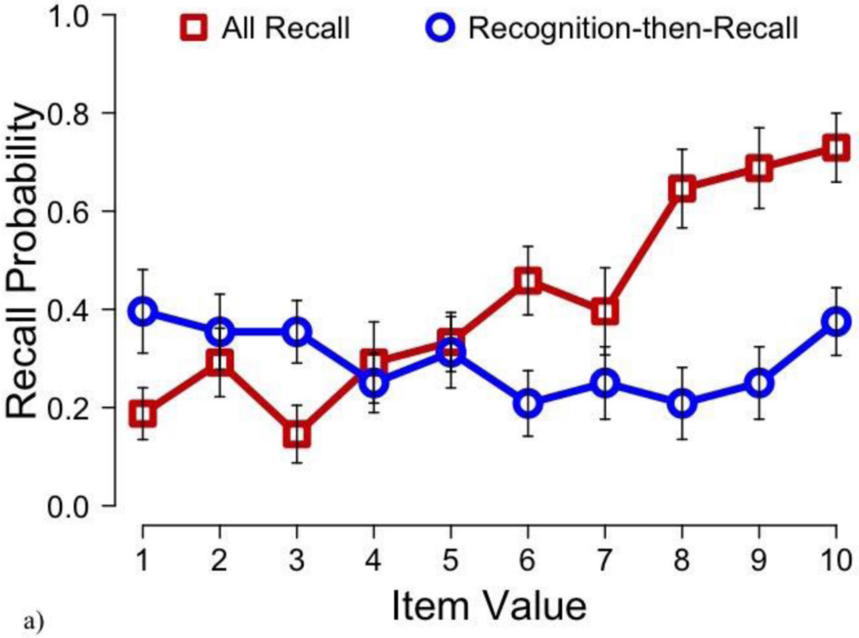
Underwood BJ. False recognition produced by implicit verbal responses. Journal of Experimental Psychology. 1965; 70:122–129. [PubMed: 14315122]

Winne, PH., Hadwin, AF. Studying as self-regulated learning. In: Hacker, DJ.Dunlosky, J., Graesser, AC., editors. Metacognition in educational theory and practice. Hillsdale, NJ: LEA; 1998. p. 277-304.

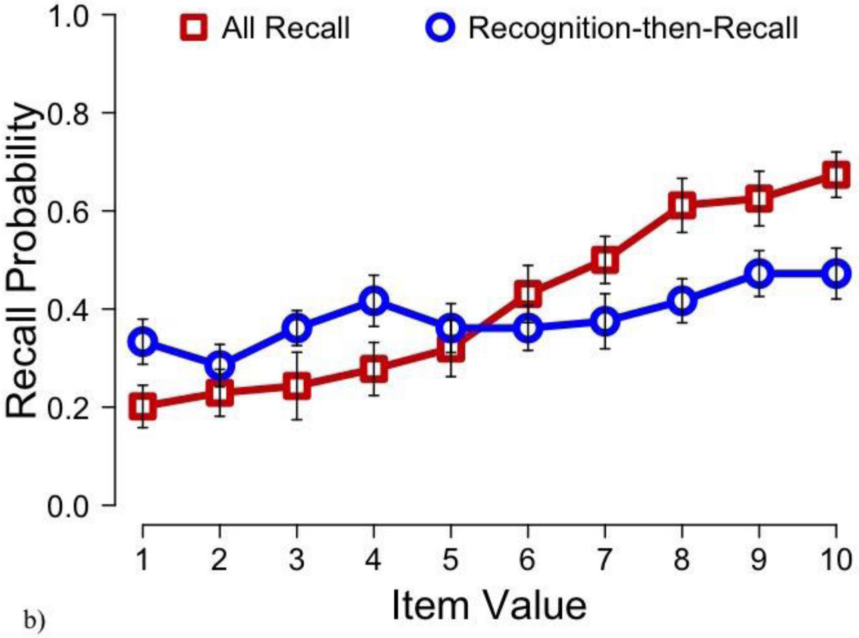
Wixted JT. Dual-process theory and signal-detection theory of recognition memory. Psychological Review. 2007; 114:152–176. [PubMed: 17227185]

Yonelinas AP. The nature of recollection and familiarity: A review of 30 years of research. Journal of Memory and Language. 2002; 46:441–517.

Yonelinas AP, Parks CM. Receiver operating characteristics (ROCs) in recognition memory: A review. Psychological Bulletin. 2007; 133:800–832. [PubMed: 17723031]


🟥 All Recall
○ Recognition-then-Recall
Recall Probability
1.0
0.8
0.6
0.4
0.2
0.0
1
2
3
4
5
6
7
8
9
10
Item Value

a)


🟥 All Recall
○ Recognition-then-Recall
Recall Probability
1.0
0.8
0.6
0.4
0.2
0.0
1
2
3
4
5
6
7
8
9
10
Item Value

b)

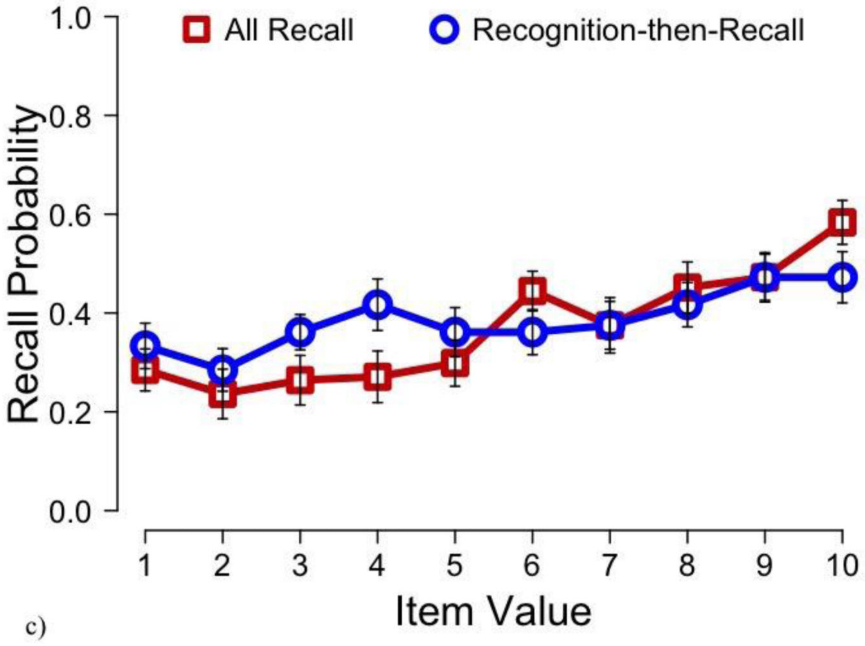
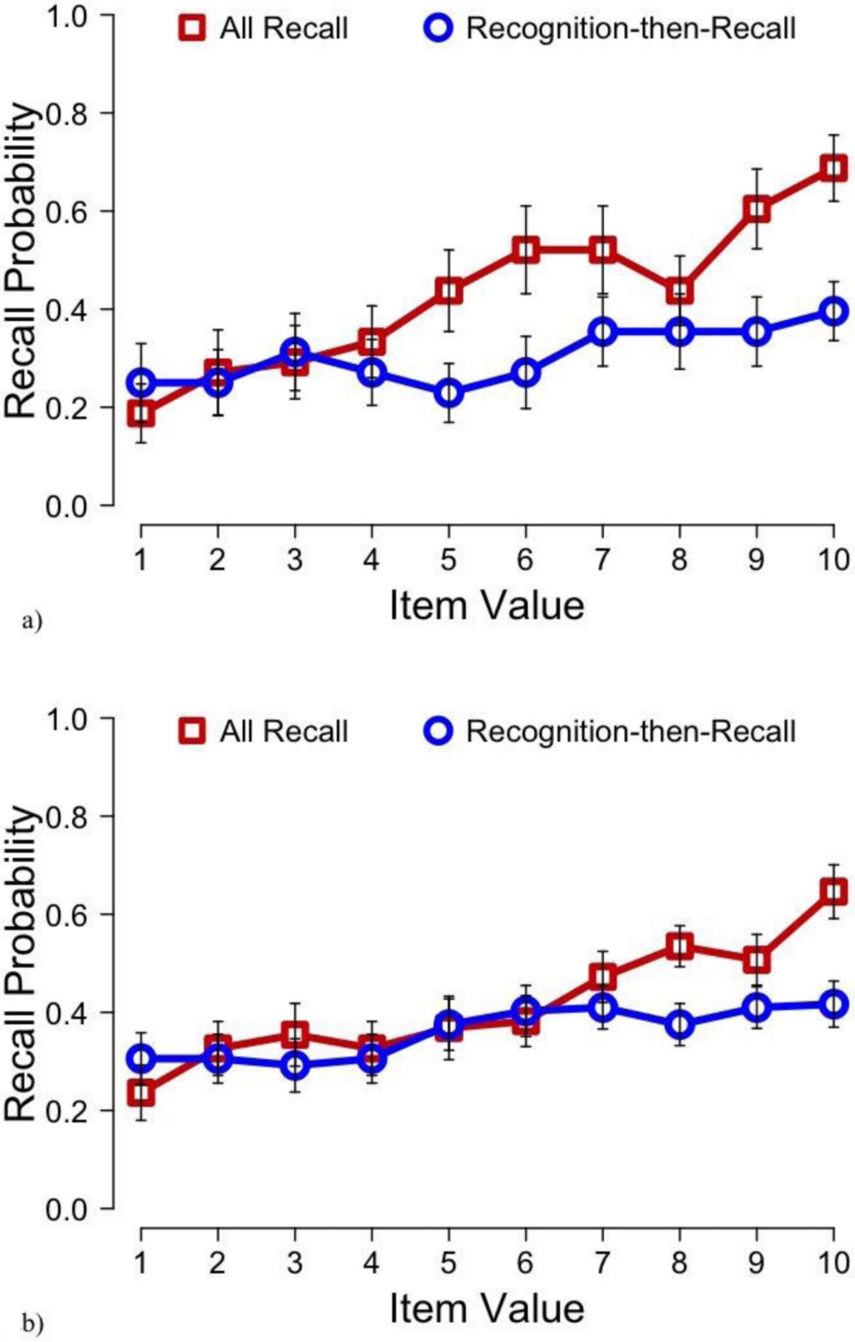


Figure 1.

Recall probability in Experiment 1 as a function of condition and item value in (a) List 5, (b) Lists 6–8, and (c) the first three expected recall tests, averaged across lists. Error bars represent standard error.


All Recall
Recognition-then-Recall
Recall Probability
1.0
0.8
0.6
0.4
0.2
0.0
1
2
3
4
5
6
7
8
9
10
Item Value

a)

b)

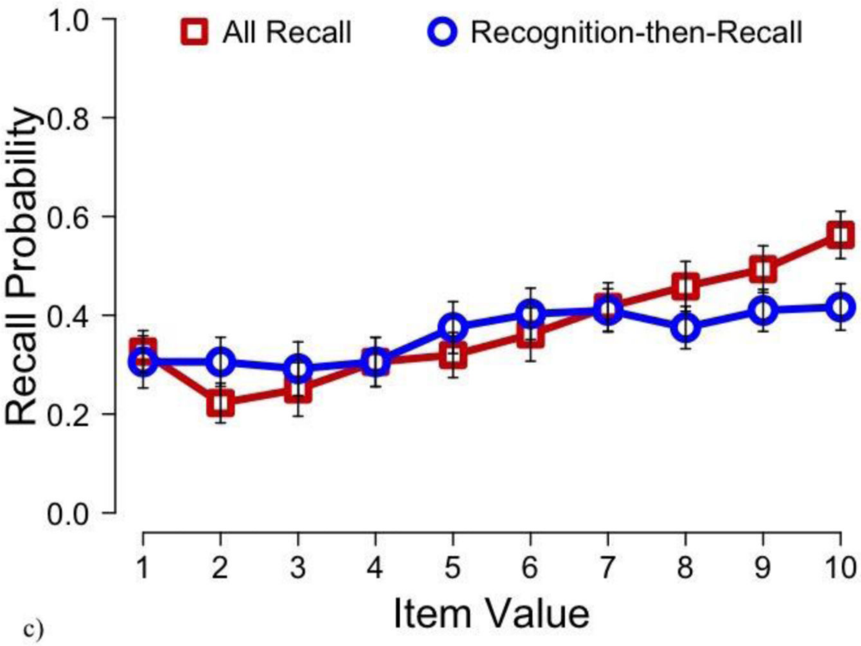


Figure 2.

Recall probability in Experiment 2 as a function of condition and item value in (a) List 5, (b) Lists 6–8, and (c) the first three expected recall tests, averaged across lists. Error bars represent standard error.

Table 1

Recall and recognition probability as a function of List and Study Condition

		Expectancy-Congruent Testing					Recall Testing				
	Condition	**L1**	**L2**	**L3**	**L4**	**Average**	**L5**	**L6**	**L7**	**L8**	**Average**
Experiment 1	All Recall (All Rc)	.38 (.12)	.36 (.14)	.36 (.11)	.39 (.15)	.37 (.10)	.42 (.16)	.43 (.15)	.40 (.13)	.41 (.20)	.42 (.14)
	Recognition-then-Recall (RgRc)	.86 (.12)	.86 (.09)	.85 (.12)	.86 (.11)	.86 (.08)	.30 (.19)	.39 (.14)	.40 (.16)	.36 (.20)	.36 (.13)
	Average						.36 (.18)	.41 (.14)	.40 (.15)	.39 (.20)	.39 (.13)
Experiment 2	All Recall (All Rc)	.34 (.15)	.40 (.20)	.36 (.14)	.41 (.15)	.38 (.11)	.43 (.15)	.44 (.17)	.40 (.19)	.41 (.17)	.42 (.14)
	Recognition-then-Recall (RgRc)	.70 (.10)	.76 (.12)	.80 (.12)	.78 (.15)	.76 (.11)	.30 (.19)	.33 (.14)	.38 (.20)	.37 (.17)	.35 (.15)
	Average						.37 (.18)	.38 (.16)	.39 (.19)	.39 (.17)	.38 (.15)

Note. “L1” through “L8” refers to Lists 1 through 8. Values for Lists 1–4 reflect recall performance for participants in the All Recall group and recognition performance for participants in the Recognition-then-Recall group. Values for Lists 5–8 reflect recall performance for all participants. Standard deviations are presented in parentheses.

Table 2

Two-level hierarchical generalized linear model of recall performance in List 5 predicted by Item Value and Study Condition

Fixed effects	**Experiment 1: List 5**	**Experiment 2: List 5**
Intercept (β_{00})	−0.43 *	−0.34 *
Predictors of intercept		
Condition (β_{01})	−0.52 *	−0.57 *
Value (β_{10})	0.31 ***	0.23 ***
Predictors of value		
Condition (β_{11})	−0.35 ***	−0.14 *
Random effects		
Intercept (r_0)	0.55 ***	0.45 ***
Value (r_1)	0.01 +	0.02 **

Note. The dependent variable is recall performance coded as 0 (*not recalled*) or 1 (*recalled*). Logit link function was used to address the binary dependent variable. Level 1 models were of the form $\eta_{ij} = \pi_{0j} + \pi_{1j}$ (Value). Level 2 models were of the form $\pi_{0j} = \beta_{00} + \beta_{01}$ (Condition) + $r_{0j}, \pi_{1j} = \beta_{10} + \beta_{11}$ (Condition) + r_{1j}. Condition was coded as 0 = *All Recall* and 1 = *Recognition-then- Recall*.

+ $p < .10$

* $p < .05$

** $p < .01$

*** $p < .001$

Table 3

Two-level hierarchical generalized linear model of recall performance predicted by Item Value, List, and Study Condition

Fixed effects	**Experiment 1: Lists 6–8**	**Experiment 1: Expected Recall Tests 1–3**	**Experiment 2: Lists 6–8**	**Experiment 2: Expected Recall Tests 1–3**
Intercept (β_{00})	−0.53**	−0.60***	−0.43*	−0.61***
Predictors of intercept				
Condition (β_{01})	−0.02	0.10	−0.22	−0.03
Value (β_{10})	0.32***	0.17***	0.19***	0.14***
Predictors of value				
Condition (β_{11})	−0.25***	−0.09*	−0.11*	−0.07
List (β_{20})	−0.10	−0.07	−0.05	0.01
Predictors of list				
Condition (β_{21})	0.01	−0.01	0.14	0.06
List × Value (β_{30})	0.05	0.02	−0.02	0.06*
Predictors of list × value				
Condition (β_{31})	−0.05	−0.02	0.03	−0.04
Random effects	**Variance**	**Variance**	**Variance**	**Variance**
Intercept (r_0)	0.41***	0.18***	0.48***	0.33***
Value (r_1)	0.10**	0.08**	0.01	0.01
List (r_2)	0.03***	0.01***	0.03***	0.01***
List × Value (r_3)	0.01+	0.004+	0.002	0.004

Note. The dependent variable is recall performance coded as 0 (*not recalled*) or 1 (*recalled*). Logit link function was used to address the binary dependent variable. Level 1 models were of the form $\eta_{ij} = \pi_{0j} + \pi_{1j}$ (Value) + π_{2j} (List) + π_{3j} (List × Value). Level 2 models were of the form $\pi_{0j} = \beta_{00} + \beta_{01}$ (Condition) + $r_{0j}, \pi_{1j} = \beta_{10} + \beta_{11}$ (Condition) + $r_{1j}, \pi_{2j} = \beta_{20} + \beta_{21}$ (Condition) + r_{2j}, $\pi_{3j} = \beta_{30} + \beta_{31}$ (Condition) + r_{3j}. Condition was coded as 0 = *All Recall* and 1 = *Recognition-then-Recall*. Note that "Expected Recall Tests 1–3" refers to the first three lists following which participants were told to expect recall tests: Lists 1–3 for participants in the All Recall condition and Lists 6–8 for participants in the Recognition-Recall condition.

+ $p < .10$

* $p < .05$

** $p < .01$

*** $p < .001$