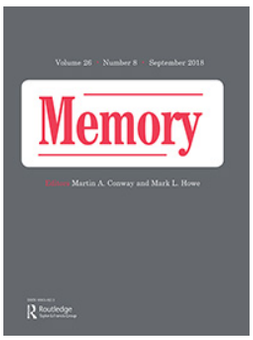

Memory

ISSN: 0965-8211 (Print) 1464-0686 (Online) Journal homepage: http://www.tandfonline.com/loi/pmem20

Evaluating the contributions of task expectancy in the testing and guessing benefits on recognition memory

Mark J. Huff, Tyler J. Yates & David A. Balota

To cite this article: Mark J. Huff, Tyler J. Yates & David A. Balota (2018) Evaluating the contributions of task expectancy in the testing and guessing benefits on recognition memory, Memory, 26:8, 1065-1083, DOI: 10.1080/09658211.2018.1467929

To link to this article: https://doi.org/10.1080/09658211.2018.1467929

Published online: 03 May 2018.

MEMORY
2018, VOL. 26, NO. 8, 1065–1083
https://doi.org/10.1080/09658211.2018.1467929

Routledge
Taylor & Francis Group

Evaluating the contributions of task expectancy in the testing and guessing benefits on recognition memory

Mark J. Huff[a], Tyler J. Yates[b] and David A. Balota[b]

[a]Department of Psychology, The University of Southern Mississippi, Hattiesburg, MS, USA; [b]Department of Psychological & Brain Sciences, Washington University, St. Louis, MO, USA

ABSTRACT

Recently, we have shown that two types of initial testing (recall of a list or guessing of critical items repeated over 12 study/test cycles) improved final recognition of related and unrelated word lists relative to restudy. These benefits were eliminated, however, when test instructions were manipulated within subjects and presented after study of each list, procedures designed to minimise expectancy of a specific type of upcoming test [Huff, Balota, & Hutchison, 2016. The costs and benefits of testing and guessing on recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *42*, 1559–1572. doi:10.1037/xlm0000269], suggesting that testing and guessing effects may be influenced by encoding strategies specific for the type of upcoming task. We follow-up these experiments by examining test-expectancy processes in guessing and testing. Testing and guessing benefits over restudy were not found when test instructions were presented either after (Experiment 1) or before (Experiment 2) a single study/task cycle was completed, nor were benefits found when instructions were presented before study/task cycles and the task was repeated three times (Experiment 3). Testing and guessing benefits emerged only when instructions were presented before a study/task cycle and the task was repeated six times (Experiments 4A and 4B). These experiments demonstrate that initial testing and guessing can produce memory benefits in recognition, but only following substantial task repetitions which likely promote task-expectancy processes.

ARTICLE HISTORY
Received 5 July 2017
Accepted 16 April 2018

KEYWORDS
Test expectancy; guessing; retrieval practice; recognition; recall

A common retrieval strategy used to enhance memory output is guessing. Guessing refers to the reporting of items in a retrieval context in which very little or no memory for an items occurrence is available or when a memory candidate is plausible given the context of a studied event. Indeed, guessing has been shown to greatly increase the number of reported memory candidates, particularly when retrieval is forced under instruction (Kelley & Sahakyan, 2003; Koriat & Goldsmith, 1996; Pansky, Goldsmith, Koriat, & Pearlman-Avnion, 2009; Roediger & Payne, 1985) and when the study materials lend themselves to successful guessing through associatively or schematically related study materials (Huff, Meade, & Hutchison, 2011). Although guessing increases item output, this increase often compromises overall memory accuracy by introducing errors (Huff et al., 2011; Meade & Roediger, 2006; for review, see Roediger, Wheeler, & Rajaram, 1993) and these guesses are often made with lower confidence that they are correct suggesting awareness that accuracy is compromised (Huff et al., 2011; Kelley & Sahakyan, 2003). Yet, despite this knowledge, guessing and/or reporting items with little confidence is common (Gardiner, Java, & Richardson-Klavehn, 1996; Roediger et al., 1993). Given the potential inaccuracies that arise from guessing, it is critical to evaluate how the guessing process can impact memory accuracy, particularly when guessing is anticipated or expected over successive tests. The purpose of the present experiments is to evaluate how expectancies of guessing may affect later recognition and compare these effects to expectancies that develop through successive recall tests and restudy trials.

Given the importance of generation on later memory, there is reason to expect that guessing may affect memory accuracy. Guessing involves a self-generative process in which a memory candidate is generated internally. Generation has been shown to produce memory benefits, particularly in mixed lists in which subsets of items are generated or not generated at study (Begg & Snider, 1987; Slamecka & Graf, 1978). However, because guessing often yields incorrect items and only some of these items are generated/guessed (mimicking a mixed-list design), guessing may also produce large error rates on subsequent tests. Consistent with the memory-enhancing effects of generation, previous research has indicated that, in cases in which corrective feedback is not provided, errors generated on early tests often persist into later tests at high rates (Bartlett, 1932/1967; Huff, Davis, & Meade,

CONTACT Mark J. Huff mark.huff@usm.edu Department of Psychology, The University of Southern Mississippi, Hattiesburg, MS 39406, USA

2013; Kang et al., 2011; Kay, 1955; McDermott, 1996; Tse, Balota, & Roediger, 2010). Initial errors are therefore concerning to memory researchers as these errors may plague memory accuracy over successive tests.

Despite the concern for persistent errors, researchers have shown that when corrective feedback is provided, guessing can improve accuracy. For example, Kornell, Hays, and Bjork (2009; also see Grimaldi & Karpicke, 2012; Hays, Kornell, & Bjork, 2013; Huelser & Metcalfe, 2012) had participants study weakly related cue-target word pairs. Pairs were either intact or presented with a cue word with instruction to guess the target word. The weak association between the cue-target pair made correct guessing of the target relatively difficult (correctly guessing 4–5% of the target words), yielding a high probability of generating an error. After guessing a target word, participants were immediately presented with the intact cue-target pair as corrective feedback. Thus, at study, participants always viewed intact pairs, but whether they first had to guess an associatively related target word or not varied within subjects. On a final cued-recall test, memory for the target word was greatest for pairs that required guessing of the target word relative to when pairs were only presented intact, demonstrating that guessing, even when the guessed response was likely incorrect, can facilitate cued recall. This guessing benefit has also been found under a variety of conditions including when (a) the final test was completed after both short (5 min) and long (24 h) delays (Kornell et al., 2009, Experiment 5; Yan, Yu, Garcia, & Bjork, 2014), (b) guess vs. intact word pairs were presented between subjects (Kornell et al., 2009, Experiment 6), and (c) guessing was used to learn foreign language pairs (Potts & Shanks, 2014).

Although guessing in the above capacity has shown reliable and robust benefits for cue-target pairs with corrective feedback, it has also been shown that the guessing process can enhance memory for information used to produce the guess initially (i.e., the cue word). In a separate line of work, the effects of guessing on memory have been found in list-learning paradigms in which participants are asked to guess non-presented related words that could have plausibly been presented on that list. Huff, Coane, Hutchison, Grasser, and Blais (2012; see too Coane, Huff, & Hutchison, 2016) presented participants with associative lists that converged either directly (e.g., sandwich, butter and jam) or indirectly (e.g., sub, spread and jar) on a non-presented critical word (e.g., bread). Directly converging lists were taken from the Deese/Roediger-McDermott false memory paradigm (DRM; Deese, 1959; Roediger & McDermott, 1995) whereas the indirectly converging lists were taken from mediated false memory lists in which mediators converge upon non-presented critical words. Immediately following the study of each list, participants then completed either a set of arithmetic problems, attempted to recall the list of words, or attempted to guess the critical word. This procedure was repeated for two blocks using the same task with each block containing six study/task lists and a final recognition test. The results showed that, consistent with retrieval-practice benefits (e.g., Rawson & Dunlosky, 2011; Roediger & Karpicke, 2006), correct recognition was greater following initial recall than arithmetic. Critically, however, recognition was greatest following guessing than the other task types, demonstrating that guessing may also facilitate memory. The findings of Huff et al. are important because they demonstrate that guessing can show memory-enhancing effects akin to retrieval practice and this improvement is found when corrective feedback is not available.

In Huff et al. (2012), false recognition for critical words was also affected by guessing. On DRM lists, initial guessing acted as an implied warning and reduced false recognition relative to completing an initial recall test. In fact, successful guessing of DRM critical words was 40%, which may have been sufficiently high to enhance monitoring for these words at test. For indirectly related mediated lists, however, successful guessing of critical words was only 5%, and as a result, false recognition of mediated critical words was greatest relative to the other tasks. This latter effect was termed the *ironic effect of guessing* because the identification of critical items initially should enhance monitoring against critical items at test, but this pattern was only found on DRM (vs. mediated) lists, when critical items were more likely to be guessed initially.

More recently, Huff, Balota, and Hutchison (2016) further examined the effects of initial testing and guessing on final recognition. In their paradigm, participants studied lists that were either categorically related, weakly related through ad hoc categories (e.g., things that are green, Hunt & Einstein, 1981; Van Overschelde, Rawson, & Dunlosky, 2004), or unrelated. These lists were used to follow the previous guessing experiments using strongly and weakly related DRM and mediated lists and to examine task effects when study items do not share associations. Similar to the false memory lists, the most representative items from each of the list types were withheld from the study to be used as critical words on a final recognition test. Immediately following the study of each list, participants either completed arithmetic problems, recalled the list, attempted to guess related but non-presented words, or restudied the list in a different order. Inclusion of a restudy control task provided greater convergence with retrieval-practice experiments which often compare the effects of restudy to retrieval (see Rawson & Dunlosky, 2011; Rowland, 2014, for reviews). On final recognition, correct recognition was found to be high and equal following the recall testing and guessing tasks relative to the arithmetic and restudy tasks and these benefits were found across list types. Hence, there was both a testing and guessing effect in this study. Guessing also elevated false recognition of categorised and ad hoc lists akin to the ironic effect of guessing (Huff et al. 2012). Further, these patterns were found when the final recognition test was completed immediately after the study/task cycles were completed (Experiment 1) and when a 24-h

delay preceded the final recognition test (Experiment 2), demonstrating generalisation of the benefits of guessing.

A critical question following these experiments is *how* the guessing process affects memory for the studied list items. One possibility is that guessing operates as an implicit retrieval task where, during the guessing task, participants implicitly retrieve studied list items to generate non-presented critical words. Another possibility is that the guessing process reflects a preparatory encoding process driven by the expectancy of an upcoming task which participants employ as the study list is presented. During study, participants may deliberately try to associate the words to generate critical non-presented words. Particularly over several study/guess cycles, participants may develop an expectancy for generating critical words due to repetitions of the guessing instructions. In other words, the benefits of guessing may not reflect implicit retrieval processes that occur when the guessing task occurs but instead reflect expectancy processes that occur as the study/guess cycles are repeated over successive trials.

To explore this possibility, Huff et al. (2016; Experiment 3) manipulated task type (either guess, recall, or restudy) within subjects and presented the task instructions after each study list was presented (vs. earlier experiments which provided instructions prior to study). The rationale for this procedural change was that by presenting task instructions randomly and after study, participants would be less likely to develop an expectancy for the upcoming task which, in turn, would reduce or eliminate the effects of encoding processes. The authors reasoned that if guessing benefits are eliminated under these constraints, the effects found in previous experiments may be due to expectancy processes influencing encoding rather than implicit retrieval processes occurring at the time the guessing task was completed. Consistent with this hypothesis, Huff et al. (2016) found that the effects of guessing on correct and false recognition were eliminated under these conditions. Also noteworthy and somewhat surprising, the benefits of initial testing were similarly eliminated (i.e., recognition in the guess, testing, and restudy tasks were equivalent). Therefore, both guessing and testing effects in this paradigm may not be due to the retrieval operations during the task, but rather due to the expectancy of the upcoming task influencing the way participants are studying the lists across repeated study/task cycles. Of equal importance, these expectancy processes also appear to affect recall testing. Thus, expectancy processes, which may be affected by guessing and testing repetitions, may influence these task benefits on memory.

Previous research has shown that expectations of upcoming memory tests can qualitatively affect memory performance. Neely and Balota (1981; Balota & Neely, 1980) established expectancies of upcoming recall and recognition tests by having participants repeatedly practice study lists that were immediately followed by either a recall or recognition test separately over three repeated study/test trials. After these practice trials, participants would then complete a final study-recall/recognition trial in which the test completed either matched or mismatched their practice test type. Regardless of the match between practice and final trials, results showed that developing expectancies for an upcoming recall test lead to an increase in the processing of study items relative to an expectancy for an upcoming recognition test. Further, Szpunar, McDermott, and Roediger (2007) showed that retrieval-practice benefits were larger when participants completed five study/test cycles and expected a final cumulative test, compared to when they did not. Finally, Pu and Tse (2014) showed that repetitive testing in reported recollection metamemory judgements on a final test, suggesting that the expectancy of upcoming tests can improve the quality of an available memory representation. Collectively, it appears that over repeated study-test trials, participants may flexibly adapt their processing during the study in anticipation of the demands of a specific type of retrieval task. Hence, the benefits of testing and guessing in multiple study test cycles in recognition memory may be influenced by task-specific encoding strategies as opposed to processes only in operation at retrieval.

Although task-expectancy effects may be powerful memory modifiers, it is also unclear how task-expectancy processes develop. Specifically, is having knowledge of the upcoming task through simple instructions sufficient to induce task-expectancy effects? Or does the task need to be repeated (i.e., practiced) over several successive trials for expectancy effects to develop (cf. Neely & Balota, 1981; Szpunar et al., 2007)? The purpose of the current set of experiments is to evaluate the boundaries of both testing and guessing benefits to determine (a) factors that affect whether expectancy effects benefit memory through task instructions and/or task repetitions (i.e., what is sufficient to illicit a testing and/or guessing-expectancy effect) and (b) whether the magnitude of expectancy effects on memory may be sensitive to the number of study-task repetitions completed.

To provide a comprehensive examination of task-expectancy effects on memory, we first conducted a replication of Huff et al.'s (2016) Experiment 3 in which task type varied within subjects and task instructions were presented only after each list was studied – methods designed to eliminate memorial benefits of testing and guessing tasks. Participants were presented with categorically related, ad hoc, and unrelated lists which were then followed by either restudy, test, or guessing tasks. One of the three task instructions was then presented randomly after study. After several study-task trials, participants completed a final recognition test to assess task effects on recognition performance. Given the robust benefits of testing effects and the past evidence of benefits of guessing, we first sought to replicate and extend the Huff et al. (2016) elimination of the recall testing and guessing benefits on a later recognition test.

The remaining experiments were designed to systematically evaluate conditions that encourage the likelihood that task-expectancy effects would be engaged. In Experiment 2, participants were presented with guess, test, or restudy task instructions before the study of each list to engage task-preparatory processes with no one task occurring on consecutive lists. Experiment 3 followed this same procedure with instructions before study with the exception that study/task cycles were blocked together such that participants would repeat the task three times before completing a different task to boost expectancy through repetitions. Finally, Experiment 4 increased the number of task repetitions to six study/task cycles. Thus, Experiments 2–4 were theoretically motivated to determine whether knowledge of an upcoming task through instructions would be sufficient to induce testing and guessing benefits over restudy, or whether increasing guessing and testing repetitions would be needed to produce memory benefits.

Experiment 1: Task instructions presented after study

In Experiment 1, the effects of interpolated restudy, testing, and guessing tasks on final recognition performance were compared. Importantly, instructions for a specific list were not provided until after a list was studied so that task-expectancy processes would be mitigated at study. The purpose of this experiment was to provide a close replication of Huff et al. (2016, Experiment 3) which used an online participant sample, by extending this study to an in-lab sample and with a shorter retention interval. Following Huff et al., we expected that testing, restudy, and guessing would produce similar levels of final recognition performance.

Methods

Participants

Forty-six English-proficient participants with normal or corrected-to-normal vision were recruited from the Undergraduate Psychology Participant Pool at Washington University in St. Louis. Data from two participants were removed due to a failure to comply with task instructions, leaving 44 participants available for analyses. Mean participant age was 19.11 years ($SD = 1.26$; Range = 17–22) and mean years of formal education was 13.02 years ($SD = 1.15$; Range = 12–15). Participants were compensated with Psychology undergraduate course credit.

Materials

Study Lists. Nine lists of 20 words were used and taken from Huff et al. (2016; Experiment 3) which included three lists from strongly related categories (birds, fruits, and spices), three lists from weakly related ad hoc categories (things made of wood, things that are black, and liquids), and three lists of unrelated items. Categorised and ad hoc lists were originally taken from the Battig and Montague (1969) and Van Overschelde et al. (2004) norms. Categorised list items were longer in letter length and occurred more frequently in language in the Hyper Analogue to Language database (Lund & Burgess, 1996) using the English Lexicon Project (Balota et al., 2007), $ts > 2.31$, $ps < .01$. Categorised and ad hoc lists were similar in concreteness and familiarity in the MRC Psycholinguistic Database (Coltheart, 1981), $ts < 1.70$, $ps > .10$. Each of the three unrelated word lists consisted of 20 randomly selected words from the MRC database (Coltheart, 1981), and matched word length, frequency, concreteness, and familiarity to ad hoc lists, $ts > 1.96$, $ps > .05$. Unrelated words were not members of the categories used for the categorised or ad hoc lists.

As was done in Huff et al. (2016), the top 5 most common exemplars from each category (or 5 random items in unrelated lists) were not studied and served as critical items, and the remaining 20 exemplars served as studied list items (see Huff & Bodner, 2014, for a similar procedure). The item order on each list was randomised anew for each participant. Three separate versions were created to counterbalance task type and list type ordering effects. All versions used one categorical list, one ad hoc list, and one unrelated list for each of the restudy, recall, and guessing tasks and were counterbalanced such that each list appeared with a different task (e.g., the fruit list was presented as a restudy task in version A, a guessing task in version B, and a recall task in version C). The counterbalances were ordered such that no one task type appeared across consecutive lists.

Recognition Test. A recognition test was constructed using 270 words. Of these 270 words, 90 were words that had appeared in the study lists (10 randomly selected words from each list), 90 were control words from categorized, ad hoc, and unrelated lists that had not been studied, 45 critical items taken from the nine study lists were included, and 45 critical items taken from the non-studied control lists. The 270 words were presented in a newly randomised order for each participant.

Procedure

The procedure was a replication of Huff et al. (2016; Experiment 3) with the exceptions that an in-lab sample was used and the retention interval was 20 min (vs. 24 h). After providing consent, participants completed a demographics questionnaire that assessed participant age and educational background. Participants were then randomly assigned to one of the three different counterbalanced versions. All participants were then presented with an instruction screen that informed them that they would study lists of 20 words in preparation for a final undisclosed memory test and that each word would be displayed for 3 s. Following each list, they would then complete one of three different tasks. In the *restudy task*, participants were told that they would restudy the list of words but in a different randomised order. In the *recall task*, participants were told that

following each list presentation that they would freely recall the list of words on the computer screen by writing them down on a recall sheet for 60 s. In the *guessing task*, participants were told that five non-presented words were related to the study list on some dimension and that their task would be to try to guess what these words might be after the list was presented for 60 s. Further, participants were also told that they were required to provide at least 1 guess and that if they finished before 60 s that they could advance to the next list. All three instructions were presented on the computer screen simultaneously and participants were further told that task instructions would be presented following each list and would be random such that it would be unknown as to which task would be completed until after each list was studied. Following these general instructions, participants were randomly assigned to one of three versions to counterbalance task order and then studied nine word lists with an interpolated task immediately following each list. Words were presented in a large 48-pt sans serif font on the centre of the screen in all caps.

Following the ninth study list/interpolated task, participants completed a series of tasks for an unrelated experiment which took approximately 20 min to complete, followed by an old/new recognition test. On this test, a word was presented on the centre of the computer screen (in the same font/size from study) and participants were instructed to respond "O" for old or "N" for new using two labelled keyboard keys. Participants were instructed to press "old" if the word was studied on any of the nine lists that were studied previously and "new" if the word was not studied previously. The word remained on the computer screen until participants made a response. Participants were also encouraged to respond quickly but without compromising accuracy.

Following the completion of the recognition test, participants were fully debriefed about the nature of the experiment and the experimenter addressed any remaining questions. The experimental procedure, including the retention interval, lasted approximately 1 h.

Results

A $p < .05$ significance level was used unless otherwise noted. Partial-eta squared (η_p^2) and Cohen's *d* effect size indices were reported for all significant effects from analyses of variance (ANOVAs) and *t*-tests, respectively. Table 1 reports mean proportions of correct and false recall on the interpolated recall test and proportions of correctly guessed critical items on the guessing task. We further tested our comparisons using a Bayesian estimate of the strength supporting the null hypothesis (Masson, 2011; Wagenmakers, 2007). This analysis compares two models, one that assumes an effect and another that assumes a null effect. This analysis yields a probability estimate that the null differences are retained – a *p*-value termed p_{BIC} (Bayesian Information Criterion), thus p_{BIC} is particularly useful for highlighting the reliability of null effects. Therefore, all null effects reported are supplemented with an additional p_{BIC} analysis.

Interpolated recall

Separate one-way ANOVAs were performed on proportions of correctly recall list items and falsely recalled critical item intrusions. Correct recall was found to vary as a function of list type, $F(2, 86) = 35.34$, $MSE = .01$, $\eta_p^2 = .45$. Follow-up comparisons showed that correct recall was greater on categorised than both ad hoc lists (.52 vs. .46), $t(43) = 2.66$, $SEM = .02$, $d = 0.50$, and unrelated lists (.52 vs. .32), $t(43) = 9.33$, $SEM = .02$, $d = 1.59$. Correct recall was also higher on ad hoc than unrelated lists (.46 vs. .32), $t(43) = 4.99$, $SEM = .03$, $d = 1.00$. Thus, as expected, lists that

Table 1. Mean (SD) initial recall proportions for list items and critical items and correct guessing of critical items of categorised, ad hoc, and unrelated lists in Experiments 1–4.

	Interpolated task group	
	Recall	Guess
Experiment 1		
List items		
Categorised	.52 (.12)	–
Ad Hoc	.46 (.14)	–
Unrelated	.32 (.15)	–
Critical items		
Categorised	.05 (.11)	.25 (.21)
Ad Hoc	.02 (.10)	.08 (.14)
Unrelated	.00 (.00)	.00 (.00)
Experiment 2		
List items		
Categorised	.58 (.11)	–
Ad Hoc	.50 (.13)	–
Unrelated	.45 (.18)	–
Critical items		
Categorised	.03 (.07)	.26 (.24)
Ad Hoc	.00 (.02)	.11 (.19)
Unrelated	.00 (.00)	.00 (.00)
Experiment 3		
List items		
Categorised	.51 (.14)	–
Ad Hoc	.52 (.17)	–
Unrelated	.43 (.18)	–
Critical items		
Categorised	.03 (.08)	.26 (.19)
Ad Hoc	.00 (.03)	.16 (.23)
Unrelated	.00 (.00)	.00 (.00)
Experiment 4A		
List items		
Categorised	.48 (.14)	–
Ad Hoc	.50 (.16)	–
Unrelated	.38 (.15)	–
Critical items		
Categorised	.01 (.04)	.24 (.21)
Ad Hoc	.00 (.02)	.09 (.12)
Unrelated	.00 (.00)	.00 (.01)
Experiment 4B		
List items		
Categorised	.48 (.14)	–
Ad Hoc	.50 (.14)	–
Unrelated	.35 (.15)	–
Critical items		
Categorised	.02 (.05)	.24 (.11)
Ad Hoc	.00 (.02)	.11 (.12)
Unrelated	.00 (.00)	.00 (.00)

provided increasing levels of semantic association also produced increasing rates of correct recall.

False recall was similarly found to differ across list types, $F(2, 86) = 5.19$. $MSE = .01$, $\eta_p^2 = .11$. Critical items were falsely recalled more frequently on categorised than unrelated lists, (.05 vs. .00), $t(43) = 3.33$, $SEM = .02$, $d = 0.71$, though only numerically greater in categorised than ad hoc lists (.05 vs. .02), $t(43) = 1.64$, $SEM = .01$, $p = .11$, $p_{BIC} = .64$. Similarly, false recall did not differ between ad hoc and unrelated lists (.02 vs. .00), $t(43) = 1.53$, $SEM = .01$, $p = .13$, $p_{BIC} = .68$. Increasing semantic associations therefore increased critical item false recall, though false recall was low across list types.

Interpolated guessing

Proportions of correctly guessed critical items for guess lists were computed by taking the total number of correctly guessed items divided by the total possible (five critical items per list) and were analysed as in recall. Correctly guessed items were found to differ across list types, $F(2, 86) = 36.73$, $MSE = .02$, $\eta_p^2 = .46$, with correct guessing greater on categorised lists than ad hoc (.25 vs. .08), $t(43) = 4.88$, $SEM = .04$, $d = 0.99$, and unrelated lists (.25 vs. .00), $t(43) = 7.93$, $SEM = .03$, $d = 1.69$. Guessing was also greater on ad hoc than unrelated lists (.08 vs. .00), $t(43) = 3.72$, $SEM = .02$, $d = 0.79$. Therefore, as in correct recall, stronger semantic relations in each list type increased correct guessing rates.

The amount of time that participants spent guessing critical items following study was also tabulated. On average, participants reported spending 49.46 s guessing the critical non-presented items, which is slightly lower than the amount of time granted on the recall task to complete the test and on the restudy task to review the list. Due to a computer error, guessing time was not available for one participant.

Recognition

Table 2 displays proportions of "old" responses to studied list items, non-studied list item controls, and critical items from studied and non-studied lists. Means were separated for each of the interpolated tasks and list types. Recognition proportions were then adjusted using a hits minus false alarms correction for correct recognition (hits for studied list items minus false alarms for non-studied list items) and for false recognition (hits for critical items minus false alarms for non-studied critical items).

A 3 (List Type) × 3 (Interpolated Task) within-subjects ANOVA was first used to compare the effects of list and task types on correct recognition of list items. Similar to results found in interpolated correct recall, correct recognition was found to differ across list types, $F(2, 86) = 24.74$, $MSE = .03$, $\eta_p^2 = .37$. Correct recognition was equivalent between categorised and ad hoc lists (.65 vs. .64), $t < 1$, $p_{BIC} = .86$, but greater on categorised than unrelated lists (.65 vs. .51), $t(43) = 5.82$, $SEM = .02$, $d = 1.01$, and on ad hoc than unrelated lists (.64 vs. .51), $t(43) = 6.06$, $SEM = .02$, $d = 0.98$.

In contrast to Huff et al.'s (2016) Experiment 3, correct recognition was found to differ across interpolated tasks, $F(2, 86) = 12.08$, $MSE = 0.02$, $\eta_p^2 = 0.22$. Interestingly, however, follow-up tests revealed greater recognition following interpolated restudy than the recall (.65 vs. .59), $t(43) = 3.02$, $SEM = 0.02$, $d = 0.41$, and guessing tasks (.65 vs. .56), $t(43) = 4.99$, $SEM = 0.02$, $d = 0.72$ – patterns opposite to that of the standard retrieval-practice effect. Correct recognition following interpolated recall was only marginally greater than that of guessing (.59 vs. .56), $t(43) = 1.80$, $p = .08$, $p_{BIC} = .58$. The Interpolated Task × List Type interaction was not reliable, $F(2, 172) = 1.22$, $MSE = 0.02$, $p = .30$, $p_{BIC} = .99$.

Analyses for proportions of critical item false recognition were completed as in correct recognition. False recognition differed across list types, $F(2, 86) = 64.54$, $MSE = 0.07$, $\eta_p^2 = 0.60$, with false recognition greater on categorised than both ad hoc (.29 vs. .17), $t(43) = 3.45$, $SEM = 0.03$, $d = 0.65$, and unrelated lists (.29 vs. −.08), $t(43) = 9.66$, $SEM = 0.04$, $d = 2.14$, and greater on ad hoc than unrelated lists (.17 vs. −.08), $t(43) = 9.71$, $SEM = 0.03$, $d = 2.17$. Thus, critical item false recognition was driven by the semantic similarity.

False recognition did not differ across interpolated task types, $F(2, 86) = 0.44$, $MSE = 0.05$, $p = .64$, $p_{BIC} = .97$, though a significant Interpolated Task × List Type interaction was found, $F(4, 172) = 2.71$, $MSE = 0.03$, $\eta_p^2 = 0.06$. The interaction reflected no task differences on unrelated and ad hoc lists, $ts < 1$, $p_{BIC}s > .72$ but on categorised lists, false

Table 2. Mean (SD) recognition proportions for list items and critical items for categorised, ad hoc, and unrelated lists as a function of interpolated task type in Experiment 1(task instructions presented after study of each list).

	Interpolated Task		
	Restudy	Recall	Guess
List items			
Categorised	.90 (.09)	.88 (.13)	.85 (.11)
Controls		.23 (.11)	
Ad Hoc	.85 (.16)	.78 (.14)	.72 (.19)
Controls		.14 (.11)	
Unrelated	.72 (.18)	.65 (.21)	.65 (.19)
Controls		.16 (.10)	
Corrected recognition			
Categorised	**.67 (.18)**	**.65 (.18)**	**.63 (.13)**
Ad Hoc	**.71 (.17)**	**.64 (.16)**	**.58 (.18)**
Unrelated	**.56 (.20)**	**.49 (.21)**	**.48 (.18)**
Task average	**.65 (.12)**	**.59 (.13)**	**.56 (.11)**
Critical items			
Categorised	.49 (.31)	.44 (.30)	.37 (.28)
Controls		.15 (.13)	
Ad Hoc	.28 (.23)	.34 (.23)	.34 (.25)
Controls		.15 (.13)	
Unrelated	.13 (.16)	.10 (.14)	.11 (.15)
Controls		.19 (.13)	
Corrected recognition			
Categorised	**.34 (.30)**	**.30 (.28)**	**.23 (.30)**
Ad Hoc	**.13 (.20)**	**.19 (.19)**	**.19 (.23)**
Unrelated	**−.06 (.15)**	**−.09 (.15)**	**−.08 (.14)**
Task Average	**.14 (.13)**	**.13 (.14)**	**.11 (.15)**

Note: $N = 44$. Boldface indicates means used in analyses.

recognition was significantly lower between the guess and restudy tasks (.23 vs. .34), $t(43) = 1.99$, $SEM = .04$, $p = .05$, $d = 0.42$, but equivalent between the guess and recall tasks (.23 vs. .30), $t(43) = 1.39$, $SEM = .04$, $p = .17$, $p_{BIC} = .72$, and the restudy and recall tasks (.34 vs. .30), $t < 1$, $p_{BIC} = .80$. This interaction suggests that the lower false recognition when attempting to guess non-presented critical items from categorised lists may have encouraged participants to detect false items and subsequently reject them on a final test. When lists were unrelated or weakly unrelated to critical items, guessing was less successful and did not reduce later false recognition. Thus, guessing appears to have reduced false recognition, but only when those critical items were strong in association to other list items and hence were likely to be identified.

Discussion

The purpose of Experiment 1 was to compare recall testing and guessing interpolated tasks on final recognition relative to a restudy control task as a replication of Huff et al. (2016; Experiment 3). Task instructions were withheld until after study of each list and no tasks occurred on consecutive lists as a means of reducing task-expectancy effects. Consistent with Huff et al., we eliminated the benefits of both recall testing and guessing, but in departure, restudy now showed improved correct recognition over testing and guessing tasks. Of course, greater recognition following restudy may be due to methodological differences employed in our current study, particularly the short retention interval (20 min vs. 24 h in the original Huff et al. study) which has shown restudy benefits over testing in other studies (e.g., Roediger & Karpicke, 2006). Despite this difference, however, the pattern remains consistent with our prediction that efforts to reduce task expectancies can mitigate recognition gains typically found with interpolated testing and guessing tasks.

Turning to non-presented critical items, categorised lists produced the greatest false recognition rates followed by ad hoc and unrelated lists, but task differences were found on categorised lists where interpolated guessing reduced false recognition relative to interpolated restudy. This pattern likely reflects detection and rejection of critical lures at test through successful guessing. Since correct guessing was greatest on categorised lists, it is reasonable that rejection of these incorrect items at test would be most likely given they were self-generated during the guessing task relative to the ad hoc and unrelated lists where correct guessing was very rare.

Experiment 2: Task instructions presented before study

Since Experiment 1 demonstrated that a method to restrict task expectancies also eliminated guessing and recall benefits over restudy, the remaining experiments examined whether methods designed to enhance expectancy processes would produce guessing and testing benefits. In Experiment 2, we first evaluated whether simply instructing participants of an upcoming task before study would be sufficient to produce testing and guessing benefits on final recognition. If participants are aware of the type of task that they will complete before study, they may qualitatively alter their processing of the study list in preparation for the task which may affect how well the list is remembered. To examine this possibility, prior to the presentation of each list, participants were presented with instructions specifying the task that they will complete immediately following study of the list. If recall and guessing benefits in correct recognition are found using this procedure, then knowledge of the upcoming task may shape encoding in the absence of any explicit study strategies. Thus, the primary difference in Experiment 2 is the timing of the task instructions (i.e., the instructions are presented immediately before versus after the study of each list.

Methods

Participants

Forty-eight English-proficient participants with normal or corrected-to-normal vision were recruited from the Undergraduate Psychology Participant Pool at Washington University. Data from one participant were eliminated due to failure to complete the experiment leaving 47 participants available for analyses. Mean participant age was 19.19 years ($SD = 1.61$; Range = 17–21) and mean years of formal education was 13.04 years ($SD = 1.15$; Range = 12–16). Participants were compensated with Psychology course credit.

Materials and procedure

The only change in this experiment is that prior to each study list, participants were provided with task instructions and told that they would complete the task after the upcoming list was studied. Restudy, recall, and guessing task instructions were identical to those used in Experiment 1. Study tasks were again counterbalanced such that no one study task was completed on consecutive lists

Results

Table 1 reports the mean proportions of correct recall and false recall on the interpolated recall test and the proportion of correctly guessed critical items on the guessing task. Data were analysed as in Experiment 1.

Interpolated recall

Correct recall was found to vary as a function of list type, $F(2, 92) = 14.68$, $MSE = .02$, $\eta_p^2 = .24$. Similar to Experiment 1, recall was significantly greater on categorised than ad hoc lists (.58 vs. .50), $t(46) = 3.61$, $SEM = .02$, $d = 0.65$, and unrelated lists (.58 vs. .45), $t(46) = 4.87$, $SEM = .03$, $d = 0.89$. Correct recall was also greater on ad hoc than unrelated

lists (.50 vs. .45), $t(46) = 2.25$, $SEM = 0.03$, $d = 0.36$, showing a clear positive relationship with semantic relatedness.

False recall was again quite low but found to differ across list types, $F(2, 92) = 4.79$, $MSE = .01$, $\eta_p^2 = 0.10$. False recall was marginally greater on categorised than ad hoc lists (.03 vs. .00), $t(46) = 1.95$, $SEM = .01$, $p = .06$, $d = 0.42$, but significantly higher on categorised than unrelated lists (.03 vs. .00), $t(46) = 2.60$, $SEM = .01$, $d = 0.53$. False recall was equivalent between ad hoc and unrelated lists (.00 vs. .00), $t < 1$, $p_{BIC} = .87$.

Interpolated guessing

Correct guessing of non-presented critical items was also found to differ across list types, $F(2, 92) = 31.74$, $MSE = 0.03$, $\eta_p^2 = 0.41$. Correct guessing was greater on categorised lists than both ad hoc (.26 vs. .11), $t(46) = 4.31$, $SEM = .04$, $d = 0.72$, and unrelated lists (.26 vs. .00), $t(46) = 7.44$, $SEM = .03$, $d = 1.53$. Correct guessing was greater on ad hoc than unrelated lists (.11 vs. .00), $t(46) = 3.93$, $SEM = .03$, $d = 0.81$. The mean time spent guessing the critical items per list was 44.90 s.

Recognition

Table 3 displays the proportions of old responses to studied list items, critical items, and their respective control items as a function of interpolated task type. Correct recognition again differed across list types, $F(2, 92) = 18.05$, $MSE = .03$, $\eta_p^2 = 0.28$. Correct recognition was equivalent between categorised and ad hoc lists (.64 vs. .63), $t < 1$, $p_{BIC} = .85$, but greater on categorised than unrelated lists (.63 vs. .53), $t(46) = 4.23$, $SEM = .02$, $d = 0.56$, and greater on ad hoc than unrelated lists (.64 vs. .53), $t(46) = 6.27$, $SEM = .02$, $d = 0.60$. However, the main effect of task type was not reliable, $F(2, 92) = 1.40$, $MSE = .03$, $p_{BIC} = .96$, and the Interpolated Task × List Type interaction also failed to reach significance, $F(2, 184) = 1.41$, $MSE = .02$. Thus, presenting participants with task instructions prior to study, a procedure designed to encourage task-preparatory processes, was not effective in producing guessing and testingbenefits over restudy on correct recognition.

False recognition was also found to differ across list types, $F(2, 92) = 87.87$, $MSE = .06$, $\eta_p^2 = 0.66$, and was greater on categorised than both ad hoc (.31 vs. .14), $t(46) = 6.44$, $SEM = .03$, $d = 1.10$, and unrelated lists (.31 vs. −.07), $t(46) = 12.55$, $SEM = .03$, $d = 2.70$. False recognition was also greater on ad hoc than unrelated lists (.14 vs. −.07), $t(46) = 7.15$, $SEM = .03$, $d = 1.34$. False recognition did not significantly differ across interpolated tasks, $F < 1$, $p_{BIC} = .98$, and the Interpolated Task × List Type interaction was also not significant, $F < 1$, $p_{BIC} = .99$. Thus, as found in correct recognition, presenting task instructions prior to study did not modulate false recognition performance on the final recognition test.

Table 3. Mean (SD) recognition proportions for list items and critical items for categorised, ad hoc, and unrelated lists as a function of interpolated task lists in Experiment 2 (task instructions presented before study of each list).

	Interpolated task		
	Restudy	Recall	Guess
List items			
Categorised	.85 (.20)	.91 (.10)	.85 (.16)
Controls		.24 (.13)	
Ad Hoc	.78 (.19)	.79 (.18)	.82 (.19)
Controls		.15 (.12)	
Unrelated	.70 (.23)	.73 (.22)	.71 (.19)
Controls		.19 (.14)	
Corrected recognition			
Categorised	**.61 (.25)**	**.67 (.17)**	**.61 (.17)**
Ad Hoc	**.63 (.23)**	**.64 (.22)**	**.67 (.20)**
Unrelated	**.52 (.25)**	**.54 (.26)**	**.53 (.21)**
Task average	**.59 (.20)**	**.62 (.18)**	**.61 (.15)**
Critical items			
Categorised	.47 (.31)	.43 (.30)	.47 (.25)
Controls		.14 (.15)	
Ad Hoc	.32 (.28)	.30 (.25)	.34 (.30)
Controls		.17 (.14)	
Unrelated	.16 (.19)	.14 (.19)	.13 (.19)
Controls		.21 (.15)	
Corrected recognition			
Categorised	**.33 (.29)**	**.29 (.28)**	**.32 (.26)**
Ad Hoc	**.15 (.26)**	**.13 (.23)**	**.16 (.27)**
Unrelated	**−.05 (.20)**	**−.07 (.17)**	**−.08 (.18)**
Task average	**.14 (.18)**	**.11 (.16)**	**.14 (.16)**

Note: $N = 47$. Boldface indicates means used in analyses.

Discussion

The results from Experiment 2 revealed that presenting task instructions prior to study did not reveal any task differences on final recognition (i.e., there was no effect of testing or guessing compared to a restudy task). When considered with Experiment 1, this pattern is important because it demonstrates that simply presenting participants with upcoming task instructions alone is not sufficient to qualitatively alter how they process the list at study. Although the results from the present study suggest that initial instructions equate recall and guessing tasks to restudy, they do differ from Experiment 1 in which initial restudy produced a correct recognition advantage over recall and guess tasks. Given these results, it is possible that initial instruction presentations may affect processing of study lists for the recall and guess lists, but only enough to equate recall and guessing tasks to the level of, but not greater than, restudy. Although it is possible that initial task instructions may produce a slight correct recognition enhancement, Experiment 2 demonstrates that this procedure is insufficient to produce recall and guessing benefits beyond that of restudy on correct recognition, which are the critical comparisons.

A possible reason for the ineffectiveness of initial instructions to produce task benefits may be due to the lack of practice or repetition that participants received on each task set. In Huff et al. (2016; Experiments 1 and 2), participants completed six study-task cycles in each of two blocks prior to completing a final recognition test when testing and guessing benefits were found. Under these conditions, participants repeated the task over several lists in which they may have developed and refined a task strategy that qualitatively shifts how lists are

encoded during study. It is therefore possible that, in addition to instructions about the upcoming task, participants may require sufficient practice with a particular task before participants sufficiently modify their encoding strategies in anticipation of an upcoming test. Indeed, Huff et al. (2016) reported that testing and guessing benefits over restudy were significantly larger when participants had already completed six study/task cycles on the second block of trials relative to the first. Thus, the critical factor to interpolated task effects may be the amount of practice that participants complete through study/task cycle repetitions resulting in a task expectancy that may shape encoding of study lists.

Experiment 3: Task type repetition over 3 lists

In Experiment 3, we sought to examine whether repetitions of study/task cycles coupled with knowledge of the upcoming task are sufficient to produce testing and guessing benefits over restudy. To maintain consistency with the previous experiments, we manipulated task type within subjects by blocking restudy, recall, and guessing tasks together over three consecutive lists. If it is necessary to receive the same task multiple times to observe expectancy effects, then blocking study lists and tasks sequentially should produce an advantage for recall and guessing tasks over restudy.

Methods

Participants

Fifty English-proficient participants with normal or corrected-to-normal vision were recruited from the Undergraduate Psychology Participant Pool at Washington University. Data from one participant were eliminated due to failure to comply with experiment instructions leaving 49 participants available for analyses. Mean participant age was 19.68 years ($SD = 1.36$; Range = 18–23) and mean years of formal education was 13.45 years ($SD = 1.21$; Range = 12–15). Participants were compensated with research credit for an undergraduate Psychology course.

Materials and procedure

All materials and procedures in Experiment 3 were identical to that of Experiment 2 with the exception that study lists were blocked by task type as a means to increase task-expectancy processes through task repetition. Three study blocks were created in which one of the three tasks (restudy, recall, or guessing) was completed. Task instructions were always provided prior to the study of each block to further enhance task expectancy. Each study block contained one categorised, ad hoc, and unrelated list type. The order of study tasks and the list types within each block were counterbalanced across participants. The final recognition test was completed after all three blocks and their interpolated tasks were completed with a 20-min retention interval separating the final task block and the recognition test.

Results

Table 1 reports the mean proportions of correct recall and false recall on the interpolated recall test and the proportion of correctly guessed critical items on the guessing task. Data analyses were similar to Experiment 1.

Interpolated recall

Correct recall was found to differ across list types, $F(2, 96) = 6.11$, $MSE = .02$, $\eta_p^2 = 0.11$. Recall was equivalent between categorised and ad hoc lists (.51 vs. .52), $t < 1$, $p_{BIC} = .87$, but significantly greater on categorised than unrelated lists (.51 vs. .43), $t(48) = 2.65$, $SEM = .03$, $d = 0.46$, and greater on ad hoc than unrelated lists (.52 vs. .43), $t(48) = 3.05$, $SEM = 0.3$, $d = 0.49$.

False recall was again found to be significantly different across list types, $F(2, 96) = 4.62$, $MSE = .01$, $\eta_p^2 = .09$, though again at floor. False recall was marginally greater on categorised than ad hoc lists (.03 vs. .00), $t(48) = 1.95$, $SEM = 0.01$, $p = .06$, $d = 0.40$, $p_{BIC} = .51$, and significantly greater than unrelated lists (.03 vs. .00), $t(48) = 2.45$, $SEM = 0.01$, $d = 0.50$. False recall for the ad hoc lists was equivalent to unrelated lists (.00 vs. .00), $t < 1$, $p_{BIC} = .88$.

Interpolated guessing

Correct guessing of non-presented critical items was also found to differ across list types, $F(2, 96) = 29.93$, $MSE = .03$, $\eta_p^2 = 0.38$. Correct guessing was greater on categorised lists than both ad hoc (.26 vs. .16), $t(48) = 2.37$, $SEM = .04$, $d = 0.56$, and unrelated lists (.26 vs. .16), $t(48) = 9.62$, $SEM = .03$, $d = 1.94$. Correct guessing was also greater on ad hoc than unrelated lists (.16 vs. .00), $t(48) = 4.90$, $SEM = .03$, $d = 0.99$. The mean time spent guessing critical items per list was 41.26 s.

Recognition

Table 4 displays the proportions of old responses to studied list items, critical items, and their respective control items as a function of list type and interpolated task. Correct recognition differed again across list types, $F(2, 96) = 11.33$, $MSE = .04$, $\eta_p^2 = 0.19$. Correct recognition was equivalent between categorised and ad hoc lists (.67 vs. .66), $t < 1$, $p_{BIC} = .85$, but significantly greater on categorised than unrelated lists (.67 vs. .57), $t(48) = 4.65$, $SEM = .02$, $d = 0.74$. Correct recognition for ad hoc lists was also greater than unrelated lists (.67 vs. .56), $t(48) = 3.87$, $SEM = .02$, $d = 0.55$.

The main effect of Task Type was right at the significance level, $F(2, 96) = 3.08$, $MSE = .03$, $p = .05$, $\eta_p^2 = .06$. Follow-up tests revealed that, in contrast to our prediction that repeated tasks would facilitate recall and guessing-expectancy benefits, restudy produced a small but reliable increase in correct recognition over guessing (.65 vs. .61), $t(48) = 2.21$, $SEM = .02$, $d = 0.30$, but not over recall (.65 vs.

Table 4. Mean (SD) recognition proportions for list items and critical items of categorised, ad hoc, and unrelated lists as a function of interpolated task lists in Experiment 3 (study lists blocked by task type across three study lists).

	Interpolated task		
	Restudy	Recall	Guess
List items			
Categorised	.92 (.09)	.87 (.14)	.85 (.17)
Controls		.20 (.10)	
Ad Hoc	.82 (.18)	.78 (.18)	.81 (.17)
Controls		.15 (.11)	
Unrelated	.74 (.23)	.75 (.17)	.68 (.22)
Controls		.15 (.10)	
Corrected recognition			
Categorised	**.72 (.15)**	**.67 (.16)**	**.64 (.17)**
Ad Hoc	**.67 (.21)**	**.64 (.21)**	**.66 (.19)**
Unrelated	**.58 (.23)**	**.59 (.20)**	**.53 (.22)**
Task average	**.65 (.16)**	**.63 (.11)**	**.61 (.15)**
Critical items			
Categorised	.45 (.30)	.45 (.33)	.48 (.26)
Controls		.12 (.14)	
Ad Hoc	.29 (.26)	.29 (.26)	.32 (.29)
Controls		.11 (.10)	
Unrelated	.13 (.17)	.13 (.17)	.12 (.19)
Controls		.18 (.14)	
Corrected recognition			
Categorised	**.34 (.30)**	**.33 (.31)**	**.37 (.25)**
Ad Hoc	**.18 (.25)**	**.18 (.25)**	**.21 (.28)**
Unrelated	**−.05 (.16)**	**−.05 (.15)**	**−.07 (.16)**
Task average	**.16 (.17)**	**.16 (.18)**	**.17 (.15)**

Notes: $N = 49$. Boldface indicates means used in analyses.

.63), $t(48) = 1.54$, $SEM = .02$, $p = .13$, $p_{BIC} = .67$. Guessing and recall were also statistically equivalent (.61 vs. .63), $t(48) = 1.16$, $SEM = .02$, $p = .25$, $p_{BIC} = .78$, and the Interpolated Task × List Type interaction was not reliable, $F(4, 192) = 1.87$, $MSE = .02$, $p = .12$, $p_{BIC} = .73$. Therefore, blocking three lists by task type as a means of encouraging task-preparatory processing did not produce a benefit for recall and guessing tasks over restudy. Instead, restudy was found to be slightly greater than guessing.

Turning to false recognition of critical items, a significant main effect of List Type was found, $F(2, 96) = 95.70$, $MSE = .07$, $\eta_p^2 = .67$. Similar to false recall, false recognition on categorised lists was greater than both ad hoc (.36 vs. .20), $t(48) = 5.23$, $SEM = .03$, $d = 0.82$, and unrelated lists (.36 vs. −.06), $t(48) = 12.72$, $SEM = .03$, $d = 2.39$. False recognition was also greater on ad hoc than unrelated lists (.20 vs. −.06), $t(48) = 9.21$, $SEM = .03$, $d = 1.89$. False recognition was not sensitive to task effects, however, as both the main effect of Interpolated Task and the Interpolated Task × List Type interaction were not significant, Fs < 1, p_{BIC}s > .85.

Discussion

Testing and guessing interpolated tasks again failed to produce correct recognition benefits over restudy when task instructions were presented before study and when tasks occurred over three successive study/task cycles. In fact, restudy was again found to be advantageous over guessing as in Experiment 1, though this effect was small and was not found when restudy was compared to interpolated recall testing. Of course, it is possible that three study/task repetitions were simply not sufficient for participants to develop a task expectancy to modify their encoding of the study lists. Previous studies that have reported task-expectancy effects have used five or more practice lists before task effects were found (e.g., Huff et al. 2012, 2016; Szpunar et al., 2007). Given this prior research, it is possible that even more repetitions may be necessary to show task effects, a boundary that we examine in Experiments 4A and 4B.

Experiment 4A: Task type repetition over 6 lists

In Experiments 4A and 4B, we evaluated whether testing and guessing task effects would occur on final recognition if participants completed a total of six study/task repetitions. Huff et al. (2012, 2016) found that both recall and guessing tasks produced recognition benefits over restudy with two blocks of six task repetitions, but it is important to note that in these experiments, task type was manipulated between subjects rather than within subjects as in the current experiments. It is therefore unclear as to whether previous testing and guessing benefits were due to the tasks themselves or the tasks in addition to a between-subjects design. Experiments 4A and 4B employed a within-subjects design of task type while increasing the number of study/task repetitions to six to determine whether recall and guessing effects on recognition could be found. We expected that blocking study/tasks into six consecutive cycles across three blocks would be a stronger method for encouraging guessing and testing expectancies.

In Experiment 4A, all participants were presented with restudy, recall, and guess instructions prior to the presentation of study lists. They then completed six study/task cycles in which the same task was repeated. After the six cycles, participants then completed a final recognition test over that block. This procedure was repeated for two additional blocks, resulting in a total of three recognition tests, each of which testedlists that were studied separately using restudy, recall, or guessing tasks. The experimental design therefore manipulated task type within subjects, but between blocks. This procedure does differ somewhat from Experiment 3 in which a single final recognition test was completed for all lists studied using the three study tasks. We implemented this change for two reasons. First, blocking recognition tests by task type allows for a closer comparison to Huff et al. (2016) who similarly used blocks containing six study lists, but with an extension to a within design. Second, utilising three smaller recognition tests vs. one large test (180 items vs. 540 items) made the length of each recognition test more similar to the test length in Huff et al. (2016; 180 items) and our previous experiments (270 items). Thus, the use of three tests is more comparable to previous experiments and reduces the likelihood of participant test fatigue by spreading out the total number of items tested.

Experiment 4B was a direct replication of 4A using a different sample. Since our previous experiments failed to

show any task differences, we sought to increase our confidence in the reliability of any task-related effects. To this end, we collected data from two samples, each of which was twice the size of the samples used in the previous experiments.

Method

Participants

One hundred twenty-two individuals were recruited using Amazon's Mechanical Turk (see Mason & Suri, 2012, for an overview).[1] All participants reported English language proficiency, resided in either Canada or the United States, and had normal or corrected-to-normal vision at the time of testing. Mean reported age was 36.34 years ($SD = 11.98$, range = 20–67) and mean years of formal education was 15.13 ($SD = 2.98$, range = 12–21). Two participants reported cheating during the experiment and were removed, leaving 120 participants for analysis.

Materials and procedure

Materials used were similar to that of Experiment 3 with the exception that the number of lists studied was increased to six categorised, six ad hoc, and six unrelated lists, to allow for blocking of six lists to increase the number of interpolated tasks that could be repeated. The same procedure was used with the exception that lists were separated into three blocks, each of which contained six total lists (two categorised, two ad hoc, and two unrelated lists). On each block, participants completed either the guessing, recall, or restudy tasks immediately after studying each list and repeated these tasks across six study/task cycles. Immediately following the sixth study/task cycle, participants then completed a recognition test for items in that study block. The recognition test contained 180 items composed of 60 studied list items (10 from each of the 6 lists from even-numbered serial positions producing 20 items from each of the categorised, ad hoc, and unrelated lists), 60 non-presented list items taken from the same positions and list types, 30 critical items from studied lists (5 from each of the 6 lists producing 10 items from each of the categorised, ad hoc, and unrelated lists), and 30 non-presented critical items taken from the same positions and list types. No retention interval was used in this experiment to reduce the total amount of time the experiment took to complete.

Following study of the first block, participants then studied the other two blocks that also contained six lists, each using one of the other interpolated task types and completed a recognition test constructed using the same distribution of items. Task instructions were always presented prior to study of the first list in each block. The block order in which task types were completed and the list orders within each block were counterbalanced across participants to control for order effects.

Following the recognition test in the third block, participants completed a demographics questionnaire and viewed debriefing information. The demographics questionnaire was like the questionnaire completed in Experiment 3 with the exception that a cheating question was added to determine whether participants misrepresented their performance during the experiment. Specifically, participants were directly asked if they wrote down list words as they were being displayed or any other technique that may have unnaturally inflated their performance during the experiment. They were asked to answer the question honestly and that they would still be compensated even if they reported that they were cheating. The experiment lasted approximately 90 min and participants were compensated $4.50 for their participation.

Results

Table 1 reports mean proportions of correct and false recall on the interpolated recall tests and proportions of correctly guessed items on the guessing task as a function of categorised, ad hoc, and unrelated lists types.

Interpolated recall

Correct recall was again found to differ across list types, $F(2, 238) = 59.88$, $MSE = .01$, $\eta_p^2 = .34$, with correct recall slightly greater on ad hoc lists than categorised lists (.50 vs. .48), $t(119) = 2.11$, $SEM = .01$, $d = 0.14$, and greater on ad hoc than unrelated lists (.50 vs. .38), $t(119) = 10.01$, $SEM = .01$, $d = 0.79$. Correct recall was similarly greater on categorised than unrelated lists (.48 vs. .38), $t(119) = 7.68$, $SEM = .01$, $d = 0.68$. Thus, in contrast to Experiments 1–3, ad hoc lists were better remembered than categorised lists, however, recall rates were relatively similar between the two and both were considerably greater than unrelated lists.

Turning to false recall of critical items, false recall was again quite low across list types, but did differ significantly, $F(2, 238) = 7.37$, $MSE = .01$, $\eta_p^2 = .06$. False recall was greater on categorised lists than both ad hoc (.01 vs. .00), $t(119) = 2.07$, $SEM = .01$, $d = 0.26$, and unrelated lists (.01 vs. .00), $t(119) = 3.42$, $SEM = .01$, $d = 0.44$, but equivalent between ad hoc and unrelated lists (.00 vs. .00), $t < 1$, $p_{BIC} = .92$.

Interpolated guessing

Correct guessing of non-presented critical items was again found to differ across list types, $F(2, 238) = 107.24$, $MSE = .02$, $\eta_p^2 = .47$. Guessing was greater on categorised than both ad hoc (.24 vs. .09), $t(119) = 8.27$, $SEM = .02$, $d = 0.88$, and unrelated lists (.24 vs. .00), $t(119) = 12.47$, $SEM = .02$, $d = 1.61$. Guessing was also greater on ad hoc than unrelated lists (.09 vs. .00), $t(119) = 8.13$, $SEM = .01$, $d = 1.04$, demonstrating again that guessing was more successful for semantically strong lists. The mean time spent guessing was 33.87 s.

Recognition

Table 5 displays proportions of old responses to studied list items, critical items, and controls as a function of list and task type. Correct recognition was found to differ across list types, $F(2, 238) = 132.40$, $MSE = .02$, $\eta_p^2 = .53$, with correct recognition greater on categorised than ad hoc lists (.70 vs. .65), $t(119) = 5.84$, $SEM = .01$, $d = 0.23$, and on categorised than unrelated lists (.70 vs. .53), $t(119) = 13.68$, $SEM = .01$, $d = 0.86$. Correct recognition was also greater on ad hoc than unrelated lists (.65 vs. .53), $t(119) = 10.92$, $SEM = .01$, $d = 0.60$.

Critically, however, a main effect of Task Type was also found, $F(2, 238) = 4.34$, $MSE = .05$, $\eta_p^2 = .04$. Consistent with our prediction that repeated study/task cycles would lead to task-related benefits on final recognition, follow-up tests revealed that completion of the guessing task lead to an increase in correct recognition over restudy (.64 vs. .60), $t(119) = 2.40$, $SEM = .02$, $d = 0.18$, and also an increase in correct recognition in recall over restudy (.64 vs. .60), $t(119) = 2.50$, $SEM = .02$, $d = 0.20$. Correct recognition was equivalent between the guess and recall tasks (.64 vs. .64), $t < 1$, $p_{BIC} = .92$.

A significant List Type × Task Type interaction was also found, $F(4, 476) = 9.07$, $MSE = .02$, $\eta_p^2 = .07$, which showed that the guess and recall benefits over restudy depended upon the different list types. Specifically, on categorised lists, guess (.68) and recall (.72) produced no benefit over restudy (.69), $ts < 1.82$, $ps > .07$, $p_{BIC}s > .68$, though recall was greater than guess (.72 vs. .68), $t(119) = 2.73$, $SEM = .02$, $d = 0.20$. On ad hoc lists, guess was only numerically greater than restudy (.65 vs. .62), $t(119) = 1.18$, $SEM = .02$, $p = .24$, $p_{BIC} = .85$, but recall was greater than restudy (.68 vs. .62), $t(119) = 2.79$, $SEM = .02$, $d = 0.24$, and equivalent to guess (.68 vs. .65), $t(119) = 1.54$, $SEM = .02$, $p = .13$, $p_{BIC} = .77$. On unrelated lists, guess produced an advantage over both restudy (.59 vs. .49), $t(119) = 4.26$, $SEM = .02$, $d = 0.43$, and recall (.59 vs. .52), $t(119) = 3.48$, $SEM = .02$, $d = 0.29$, but recall was only numerically greater than restudy (.52 vs. .49), $t(119) = 1.38$, $SEM = .02$, $p = .17$, $p_{BIC} = .81$. Thus, guess and recall benefits, though statistically reliable, were restricted to ad hoc and unrelated lists, suggesting that categorised lists may be relatively less affected by task-expectancy repetitions.

False recognition of critical items similarly showed a main effect of List Type, $F(2, 238) = 104.42$, $MSE = .05$, $\eta_p^2 = .47$, in which false recognition was greater on categorised lists than both ad hoc (.23 vs. .13), $t(119) = 6.43$, $SEM = .02$, $d = 0.61$, and unrelated lists (.23 vs. −.01), $t(119) = 12.11$, $SEM = .02$, $d = 1.56$. False recognition was also greater on ad hoc than unrelated lists (.13 vs. −.01), $t(119) = 10.16$, $SEM = .02$, $d = 1.22$. Like correct recognition, false recognition was sensitive to interpolated task types, $F(2, 238) = 32.46$, $MSE = .04$, $\eta_p^2 = .21$. False recognition was lower following the guessing task than the restudy task (.07 vs. .18), $t(119) = 7.25$, $SEM = .02$, $d = 0.94$, and lower in the recall than the restudy task (.09 vs. .18), $t(119) = 5.63$, $SEM = .02$, $d = 0.79$. False recognition was marginally lower in the guess than recall tasks (.07 vs. .09), $t(119) = 1.81$, $SEM = .01$, $p = .07$, $p_{BIC} = .68$, though false recognition in both tasks was rather low.

The effects of List Type and Task Type were qualified by a significant interaction, $F(4, 476) = 11.73$, $MSE = .04$, $\eta_p^2 = .09$. Follow-up tests revealed that false recognition was lower for both guess and recall tasks relative to restudy, but this reduction was restricted to categorised and ad hoc lists, $ts > 4.82$, $ds > 0.64$, with no task differences

Table 5. Mean (SD) final recognition proportions for list items and critical items of categorised, ad hoc, and unrelated lists as a function of interpolated task lists in Experiments 4A and 4B (study lists blocked by task type across six lists).

	Restudy			Recall			Guess		
	Categorised	Ad Hoc	Unrelated	Categorised	Ad Hoc	Unrelated	Categorised	Ad Hoc	Unrelated
Experiment 4A									
Raw recognition									
List Items	.83 (.16)	.79 (.17)	.64 (.21)	.82 (.16)	.79 (.16)	.64 (.18)	.82 (.16)	.77 (.19)	.73 (.20)
Controls	.15 (.18)	.16 (.19)	.15 (.16)	.10 (.14)	.12 (.15)	.12 (.14)	.14 (.16)	.13 (.15)	.14 (.16)
Critical Items	.50 (.31)	.41 (.27)	.14 (.16)	.34 (.26)	.23 (.19)	.13 (.16)	.34 (.26)	.24 (.21)	.14 (.17)
Controls	.16 (.20)	.18 (.20)	.16 (.19)	.16 (.18)	.13 (.15)	.12 (.16)	.18 (.20)	.19 (.22)	.15 (.19)
Corrected recognition									
List Items	**.69 (.26)**	**.62 (.25)**	**.49 (.23)**	**.72 (.21)**	**.68 (.22)**	**.52 (.23)**	**.68 (.22)**	**.65 (.26)**	**.59 (.24)**
Task Average		**.60 (.22)**			**.64 (.19)**			**.64 (.22)**	
Critical Items	**.34 (.30)**	**.23 (.26)**	**−.02 (.15)**	**.18 (.27)**	**.09 (.16)**	**.00 (.14)**	**.16 (.24)**	**.05 (.21)**	**−.01 (.14)**
Task Average		**.18 (.17)**			**.09 (.12)**			**.07 (.13)**	
Experiment 4B									
Raw recognition									
List items	.85 (.16)	.76 (.18)	.59 (.23)	.84 (.15)	.81 (.15)	.65 (.20)	.86 (.13)	.78 (.17)	.73 (.21)
Controls	.12 (.16)	.12 (.16)	.13 (.16)	.06 (.12)	.09 (.12)	.08 (.11)	.10 (.11)	.10 (.12)	.12 (.13)
Critical items	.53 (.31)	.43 (.27)	.11 (.16)	.33 (.25)	.21 (.21)	.10 (.14)	.31 (.24)	.21 (.18)	.11 (.16)
Controls	.14 (.19)	.16 (.20)	.12 (.14)	.11 (.17)	.09 (.14)	.09 (.12)	.15 (.17)	.15 (.17)	.12 (.16)
Corrected recognition									
List items	**.73 (.23)**	**.64 (.25)**	**.45 (.27)**	**.78 (.20)**	**.72 (.20)**	**.57 (.22)**	**.76 (.17)**	**.68 (.19)**	**.61 (.22)**
Task average		**.61 (.22)**			**.69 (.18)**			**.68 (.16)**	
Critical items	**.39 (.31)**	**.27 (.25)**	**−.01 (.13)**	**.22 (.28)**	**.11 (.17)**	**.02 (.12)**	**.16 (.27)**	**.06 (.19)**	**−.01 (.13)**
Task average		**.22 (.16)**			**.12 (.13)**			**.07 (.12)**	

Note: $N = 120$ in Experiment 4A; $N = 104$ in Experiment 4B. Boldface indicates means used in analyses.

found on unrelated lists, $ts < 1.15$, $ps > .25$, $p_{BIC}s > .84$. False recognition on guess and recall tasks was also equivalent across all list types, $ts < 1.53$, $ps > .12$, $p_{BIC}s > .77$. Thus, guess and recall tasks appeared to further enhance memory accuracy through a reduction in false recognition of critical items, but only when critical items were semantically related to the study list.

Experiment 4B

Participants

One hundred six individuals were recruited using Amazon's Mechanical Turk. All participants reported English language proficiency, resided in either Canada or the United States, and had normal or corrected-to-normal vision. Mean reported age was 38.12 years ($SD = 11.45$, range = 19–64) and mean years of formal education was 15.00 ($SD = 1.69$, range = 10–19). Two participants reported cheating during the experiment, leaving 104 for analysis. Data for Experiment 4B were collected 5–6 months following Experiment 4A.

Materials and procedure

All materials and procedures were identical to that of Experiment 4A.

Results

Table 1 reports mean proportions of correct and false recall on the interpolated recall tests and proportions of correctly guessed items on the guessing task as a function of categorised, ad hoc, and unrelated lists types. Data were analysed as in Experiment 4A.

Interpolated recall

Correct recall was again found to differ across list types, $F(2, 206) = 100.53$, $MSE = .01$, $\eta_p^2 = .49$, with correct recall slightly greater on ad hoc lists that categorised lists (.50 vs. .48), $t(103) = 2.31$, $SEM = .01$, $d = 0.14$, and greater on ad hoc than unrelated lists (.50 vs. .35), $t(103) = 10.87$, $SEM = .01$, $d = 0.99$. Correct recall was similarly greater on categorised than ad hoc lists (.48 vs. .35), $t(103) = 7.68$, $SEM = .01$, $d = 0.85$.

Critical item false recall was similarly quite low across list types, but did differ significantly, $F(2, 206) = 16.23$, $MSE = .01$, $\eta_p^2 = .14$. False recall was greater on categorised lists than both ad hoc (.02 vs. .00), $t(103) = 3.60$, $SEM = .05$, $d = 0.50$, and unrelated lists (.02 vs. .00), $t(103) = 4.68$, $SEM = .01$, $d = 0.65$, but equivalent between ad hoc and unrelated lists (.00 vs. .00), $t < 1$, $p_{BIC} = .92$.

Interpolated guessing

Correct guessing of non-presented critical items was again found to differ across list types, $F(2, 206) = 114.91$, $MSE = .01$, $\eta_p^2 = .53$. Guessing was greater on categorised than both ad hoc (.24 vs. .11), $t(103) = 7.90$, $SEM = .02$, $d = 0.89$, and unrelated lists (.24 vs. .00), $t(103) = 13.28$, $SEM = .02$, $d = 1.84$. Guessing was also greater on ad hoc than unrelated lists (.11 vs. .00), $t(103) = 9.02$, $SEM = .01$, $d = 1.25$, demonstrating again that guessing was more successful for semantically strong lists. The mean time spent guessing was 30.32 s.

Recognition

Table 5 displays proportions of old responses to studied list items, critical items, and controls as a function of list and task type. Correct recognition was found to differ across list types, $F(2, 206) = 154.68$, $MSE = .02$, $\eta_p^2 = .60$, with correct recognition greater on categorised than ad hoc lists (.76 vs. .68), $t(103) = 7.43$, $SEM = .01$, $d = 0.44$, and on categorised than unrelated lists (.76 vs. .54), $t(103) = 15.29$, $SEM = .01$, $d = 1.21$. Correct recognition was also greater on ad hoc than unrelated lists (.68 vs. .54), $t(103) = 11.11$, $SEM = .01$, $d = 0.74$.

Critically, the main effect of Task Type was again found, $F(2, 206) = 13.99$, $MSE = .05$, $\eta_p^2 = .12$. Consistent with our prediction that repeated study/task cycles would lead to task-related benefits on final recognition, follow-up tests revealed that completion of the guessing task lead to an increase in correct recognition over restudy (.68 vs. .61), $t(103) = 4.48$, $SEM = .02$, $d = 0.39$, and an increase in correct recognition in recall over restudy (.69 vs. .61), $t(103) = 4.37$, $SEM = .02$, $d = 0.40$. Correct recognition was equivalent between the guess and recall tasks (.68 vs. .69), $t < 1$, $p_{BIC} = .91$.

A significant List Type × Task Type interaction was also found, $F(4, 412) = 9.47$, $MSE = .01$, $\eta_p^2 = .08$, which showed that the guess and recall benefits over restudy depended upon the different list types. Specifically, on categorised lists, no differences were found between the guess and restudy tasks (.76 vs. .73), $t(103) = 1.52$, $SEM = .02$, $p = .13$, $p_{BIC} = .76$, but there was a small but reliable improvement for recall over restudy (.78 vs. .73), $t(103) = 2.10$, $SEM = .02$, $d = 0.21$, and no difference between guess and recall lists (.76 vs. .78), $t < 1$, $p_{BIC} = .87$. Consistent with Experiment 4A, on ad hoc lists, the guessing task produced a significant benefit over restudy (.68 vs. .64), $t(103) = 2.31$, $SEM = .02$, $d = 0.21$, and this benefit was also found for recall over restudy (.72 vs. .63), $t(103) = 3.99$, $SEM = .02$, $d = 0.37$. A numeric benefit was also found for recall over guessing (.72 vs. .68), but this difference was right at the level of significance, $t(103) = 1.96$, $SEM = .02$, $p = .05$, $d = 0.19$. Task benefits also extended to unrelated lists with both the guessing task producing a benefit over restudy (.61 vs. .45), $t(103) = 5.66$, $SEM = .03$, $d = 0.63$, and the recall over restudy (.57 vs. .45), $t(103) = 4.70$, $SEM = .02$, $d = 0.47$. Guessing was marginally greater than recall (.61 vs. .57), $t(103) = 1.78$, $SEM = .02$, $p = .08$, $p_{BIC} = .68$.

False recognition similarly showed a main effect of List Type, $F(2, 206) = 91.10$, $MSE = .06$, $\eta_p^2 = .47$, in which false recognition was greater on categorised lists than both ad hoc (.26 vs. .15), $t(103) = 6.41$, $SEM = .02$, $d = 0.64$, and unrelated lists (.26 vs. .00), $t(103) = 10.99$, $SEM = .02$, $d = 1.60$.

False recognition was also greater on ad hoc than unrelated lists (.15 vs. .00), $t(103) = 9.38$, $SEM = .02$, $d = 1.32$. Like correct recognition, false recognition was sensitive to interpolated task types, $F(2, 206) = 41.10$, $MSE = .04$, $\eta_p^2 = .29$. False recognition was lower in the guessing task than restudy (.07 vs. .22), $t(103) = 8.01$, $SEM = .02$, $d = 1.03$, and lower in the recall than the restudy task (.12 vs. .22), $t(103) = 5.58$, $SEM = .02$, $d = 0.66$. False recognition was lower in the guess than recall task (.07 vs. .12), $t(103) = 3.68$, $SEM = .01$, $d = 0.38$.

The effects of List Type and Task Type were qualified by a significant interaction, $F(4, 412) = 14.32$, $MSE = .03$, $\eta_p^2 = .12$. Follow-up tests revealed that false recognition was lower for both guess and recall tasks relative to restudy, and like Experiment 4A, this reduction was restricted to categorised and ad hoc lists, $ts > 4.81$, $ds > 0.64$, with no task differences found on unrelated lists, $ts < 1.51$, $ps > .13$, $p_{BIC}s > .76$. Unlike Experiment 4A, however, false recognition was lower for the guess than recall tasks on both categorised and ad hoc lists, $ts > 2.26$, $ds > 0.31$, but only numerically lower on unrelated lists, $t(103) = 1.74$, $SEM = .01$, $p = .08$, $p_{BIC} = .69$. Thus, when guessing was more successful at identifying critical items, as in the case for categorised and ad hoc lists, participants were more likely to reject these items on the recognition test.

Discussion

The primary findings of Experiments 4A and 4B were quite clear. Completing a recall or guessing task after studying a list of words had a beneficial effect on final recognition relative to restudy when instructions were presented prior to study and when tasks were completed over six consecutive study/task cycles. Specifically, in both experiments, repeated recall and guessing tasks improved correct recognition on ad hoc and unrelated lists and simultaneously decreased false recognition on categorised and ad hoc lists, an improvement to overall memory accuracy (a pattern termed a mirror effect; Glanzer & Adams, 1990). We argue that this pattern is due to task repetitions increasing expectancies and thereby promoting task-specific processing of the list during encoding. In Experiment 4B, correct recognition on categorised lists was greater following recall than restudy, but not guessing. The similarity of these benefits in two separate experiments speaks to the reliability of interpolated task benefits, but only when experimental conditions are sufficient to foster task-expectancy processes through repetition. Of additional importance, recall and guessing benefits were found using a within-subjects vs. a between-subjects design (cf. Huff et al., 2016) – a novel extension.

Although testing and guessing tasks showed recognition benefits, these benefits were sensitive to list type. For correct recognition, testing and guessing benefits were only found on ad hoc and unrelated lists and only recall showed a benefit on categorised lists in Experiment 4B, demonstrating that the effects may not be global across materials. A possible reason for this list type difference may be due to the relatively higher rate of correct recognition in general for categorised lists, relative to ad hoc and unrelated lists. Categorised lists may be less sensitive to task effects because memory representations for strongly related items are largely intact, making it difficult to further strengthen recognition of categorised items. False recognition similarly showed a task effect which was modulated by a list type effect where critical item recognition showed no task effects on unrelated lists, but the recall and guessing tasks reduced critical item recognition on categorised and ad hoc lists. This pattern is likely due categorised and ad hoc lists containing items related the critical lures. Related lists may have made the critical items more identifiable through task-expectancy processes, promoting monitoring processes at recognition. Alternatively, since correct recognition was also increased following guessing and recall tasks, these task types may have also promoted the processing of verbatim information which reduced reliance on the gist or thematic consistency of the list at test (e.g., Brainerd & Reyna, 2002). Regardless of the mechanism, however, the reduction in false recognition following the recall and guessing tasks further demonstrates that tasks, or the expectancy of upcoming tasks, improve discrimination between items that were or were not presented on the studied lists. In recall, reporting words from a studied list may induce a recall-to-reject strategy (Gallo, 2004) in which intact memory for the list can assist in rejecting non-studied lures. Whereas in guessing, participants are provided with information about the existence of critical items and asked to identify them, processes that later enhance test-based monitoring.

General discussion

The present study evaluated the role of task-expectancy processes in producing testing, guessing, and restudy interpolated task effects on correct and false recognition on categorised, ad hoc, and unrelated list types. We suggested that if participants anticipate a task that demands active processing of the study list, such as recalling a list of words from memory or attempting to guess information using a list of words, that they will modify how they encode that information in preparation for the upcoming task. Since memorial benefits of testing and guessing have been shown consistently (e.g., Huff et al., 2012, 2016; Roediger & Karpicke, 2006; Rawson & Dunlosky, 2011), it is therefore important to determine the roles anticipatory, task-specific processes play in enhancing memory. Hence, we examined specific methods designed to both reduce and enhance task-expectancy processes of testing and guessing and compare their effects on a final recognition test relative to a restudy control task. The critical finding was that methods designed to limit testing and guessing-expectancy processes did not

produce memory benefits over restudy; however, when expectancy processes were encouraged through extended task practice, both testing and guessing benefits over restudy emerged in recognition performance. To aid understanding of the task patterns across experiments, task effects collapsed across list type are plotted in Figure 1.

Experiment 1 was a close replication of Huff et al. (2016; Experiment 3) and was designed to restrict testing and guessing-expectancy processes by withholding task instructions until after a single list was studied. Under these conditions, we anticipated that because participants had no knowledge of upcoming testing and guessing tasks, they would be less likely to engage in preparatory processes at study, hampering the effectiveness of testing and guessing to improve recognition. Consistent with our hypothesis, interpolated recall and guess conditions did not increase correct recognition and were lower than that of restudy – a testing and guessing cost. The remaining experiments focused on determining those conditions that would promote expectancy effects through manipulations designed to increase awareness of an upcoming task. Our results showed that providing participants with task instructions immediately prior to study was not sufficient to induce testing and guessing recognition benefits (Experiment 2) nor was coupling initial task instructions with task repetitions over three consecutive lists (Experiment 3). Only when initial instructions were provided and the task was repeated over six study/task cycles in Experiments 4A and 4B were testing and guessing benefits over restudy found.

Our findings, particularly those showing the benefits of guessing, extend previous studies that have shown benefits of initial guessing on retention (e.g., Hays et al., 2013; Kornell et al., 2009; Potts & Shanks, 2014). As highlighted above, guessing is a common retrieval strategy, particularly when study information allows for plausible guesses. It is therefore likely that participants may expect that they will be guessing if/when memory fails or becomes exhausted, resulting in guessing-expectancy processes. Further, by including a recall test comparison, we could evaluate testing and guessing processes on final recognition. Remarkably, testing and guessing conditions produced very similar effects on final recognition although the guessing task instructed participants to generate items that were not studied, whereas in recall, list items were being directly retrieved. Thus, the results of testing and guessing are similar despite their instructions being quite different.

Collectively, our experiments provide important demonstrations regarding how specific interpolated tasks operate to produce memory benefits. First, our experiments demonstrate that testing and guessing effects are more likely to emerge when expectancy processes are high. Previously established task benefits such as retrieval-practice postulate that testing benefits emerge through the act of completing a test itself rather than through retrieval-expectancy processes such as increasing the number of available retrieval routes to facilitate the effectiveness of cues on subsequent tests (Bjork, 1975; Roediger & & Karpicke, 2006) or through the generation of implicit semantic mediators which can also be used as cues (i.e., mediator-effectiveness hypothesis, Carpenter, 2011; Pyc & Rawson, 2010). Our experiments do not question the benefits of retrieval practice, however, they do provide evidence that at least when considering the benefits of recall and guessing on later recognition performance, participants appear to develop strategies across repeated study/test trials that produce benefits over a restudy condition. These benefits may occur in two ways: First, as suggested, anticipation of an upcoming testing and guessing tasks may facilitate encoding

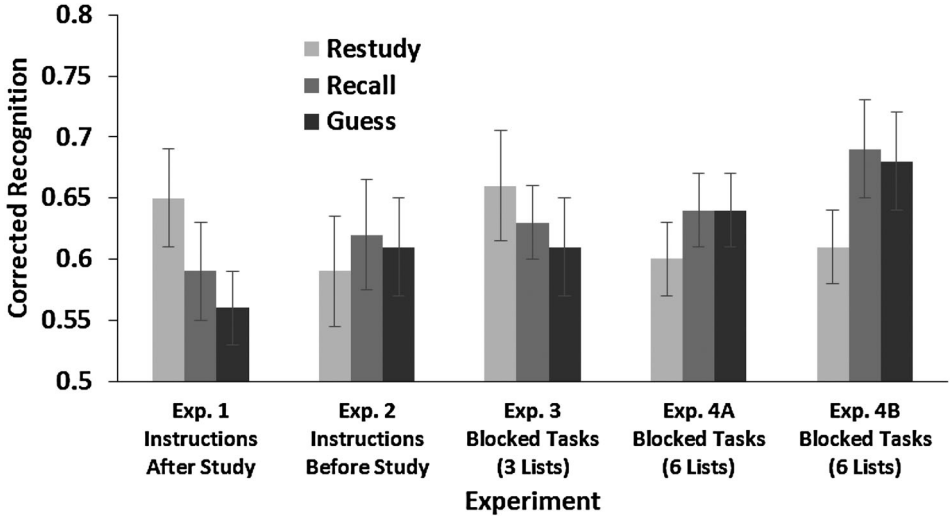

Figure 1. Mean corrected recognition data as a function of task type (collapsed across list type) from Experiments 1–4. Error bars are 95% confidence intervals for the means.

processes (cf. Balota & Neely, 1980) vs. restudy in which participants anticipate the same information again. Second, knowledge of restudy conditions may even encourage "lazy encoding" of the study lists by participants. Knowing that items will be viewed again, participants may be less engaged and loaf during the first encoding attempt. Participants completing the testing and guessing tasks, on the other hand, maybe more engaged because the items will only be presented once, while also strategically processing those items for the upcoming task.

Further, our experiments demonstrate that task-expectancy processes require sufficient task practice (at least greater than three study/task repetitions) and do not occur with simple knowledge of an upcoming task. This pattern is consistent with both Balota and Neely (1980) and Szpunar et al. (2007) who similarly showed that expectancy-type task effects were enhanced following several repetitions. Further, Huff et al. (2016) showed that when testing and guessing tasks were completed over two blocks of six lists, benefits over restudy were greater on the second block in which more task practice was completed. Together, these patterns suggest that participants may develop and then refine strategies for specific tasks, which become more effective over repetitions.

An important question regarding task expectancies is therefore how expectancy processes operate to affect memory performance. When considering the recall and guessing tasks in our experiments, we argue that it is possible that participants may engage in organisation or relational processing to improve task performance. Previous literature has suggested that the act of recall testing likely encourages organisational processing of the tested materials (Congleton & Rajaram, 2012; Zaromb & Roediger, 2010) and these processes may indeed increase with additional recall practice. Similarly, the guessing task also likely encouraged relational processing since the task instruction asked participants to associate words together in order to generate the critical missing word. Organisational and/or relational processing tasks have been shown to produce memory benefits over control tasks (Huff & Bodner, 2013). Hence, it is plausible that task expectancies may produce recognition benefits through an increase in deep relational processing. Of course, our experiments do not provide a direct index of the type or magnitude of processing that participants engage in over task repetitions. Therefore, future research should aim to determine what processes underlie task expectancies and whether these processes differ between recall and guessing tasks.

Additional evidence for the occurrence of expectancy processes may be found in the amount of time participants spent guessing critical items during the guessing task. If participants developed an expectancy for the upcoming guessing task, they may require less time during the guessing task to report guesses because they are instead generating guesses during the presentation of the list and have these guesses available when the guessing task begins. Consistent with this possibility, we note that the time spent guessing during the guessing task was greatest when expectancy processes were likely minimal as in Experiment 1 (49.46 s) and then decreased over successive manipulations to increase expectancy, such as presenting initial instructions in Experiment 2 (44.90 s), repeating tasks three times in Experiment 3 (41.26 s), and repeating tasks six times in Experiments 4A (33.87) and 4B (30.32 s). These differences were found to differ significantly, $F(4, 359) = 13.37$, $MSE = 30.67$, $\eta_p^2 = .13$, and may provide an additional index regarding the strength of task expectancies.

Although the benefits of recall over restudy are reliable after six study/task repetitions, one domain that did not show consistent task benefits was on categorised lists. This pattern differs from Huff et al. (2016) who showed that both testing and guessing benefits occurred across all list types at an equal magnitude, not just following recall as in Experiment 4B. While our data may challenge the generality of testing and guessing benefits in this paradigm across different types of materials, it is important to note that there are some methodological differences that may contribute to this divergent pattern. First, participants in Huff et al.'s (2016) experiments received considerably more task practice by repeating tasks 12 times across 2 study blocks. It is possible that these lists may simply require even more practice before task benefits emerge. Second, our manipulation of task type was also completed within subjects, but between lists, whereas Huff et al.'s (2016) tasks occurred between subjects. Design differences, which may have allowed for task carry-over effects, may have reduced task effectiveness selectively for the more memorable categorised lists. Of course, determining those factors that limit the generality of task benefits is important in future work for understanding how interpolated tasks directly benefit subsequent memory.

Further, it is important to emphasise that, although we make the argument that task-expectancy processes benefit subsequent memory, these effects have only been found on final recognition. It is unclear whether testing and guessing benefits would extend to a final recall test, which is typically believed to be a more recollection-heavy test vs. recognition, a test more sensitive to contributions of automatic (or familiarity) processes (Mandler, 1980; Yonelinas, 2002). It is therefore unclear whether guessing and testing benefits compared to restudy reflect boosts in familiarity, recollection, or some combination of the two. If the present results extend to a recall test, this would provide evidence that task effects may also enhance recollective processes since improvements on recognition could be due solely to familiarity. Simply extending the current experiments to a final recall task, however, may be challenging given the interpolated recall condition would be completing the same task again for the final test. It would therefore be necessary to dissociate those effects of interpolated recall from that of transfer-appropriate processing

(Morris, Bransford, & Franks, 1977), which would affect performance due to a test match vs. a more general task-based memory benefit. Guessing on the other hand, differs from recall, so swapping recognition for recall could be useful for determining guess and guess-expectancy benefits in final recall. Of course, no task or test is "process pure" (Jacoby, 1991; 1994), but it would be important for future work to determine those memorial processes that are augmented through task-related processes and to determine the generality of task-related benefits.

Given the preceding discussion, it should be clear that we are not proposing that existing task-related memory benefits, such as the retrieval-practice effect (Roediger & Karpicke, 2006), are entirely due to encoding-based task-expectancy effects, and likely reflect additional test-specific effects. Minimally, testing provides a restudy opportunity for information that is retrieved, and maximally produces deeper encoding of originally studied information through processes described above (e.g., Pyc & Rawson, 2010; Zaromb & Roediger, 2010). However, it is likely that expectancy processes also play a role. In many retrieval-practice experiments, participants either have knowledge prior to study that they will complete interpolated testing/restudy tasks or that they will repeat these tasks over consecutive study lists (see Rawson & Dunlosky, 2011). Even in experiments in which testing and restudy trials are interleaved, making it unclear to participants which specific task they will complete after study, this procedure yields a high number of task repetitions which, based on our data, produces the greatest task-expectancy effects. Thus, we suggest that task/test-expectancy processes may be contributing in part to these powerful memory benefits that have previously been accounted for through task-only processes.

Conclusions

The present study was conducted to evaluate how expectancies of interpolated free recall and guessing tasks can influence how a list of words is encoded and later remembered on a recognition test. The results of this study are noteworthy in that removing procedures designed to restrict expectancy processes do not produce recall testing or guessing benefits on a final recognition test, but when expectancy processes are enhanced through task repetitions, testing and guessing benefits are found. Our results also provide information regarding important boundary conditions of these expectancy processes. Testing and guessing benefits do not emerge when participants are presented with instructions about the upcoming interpolated task prior to study nor are they found when instructions are presented prior to study and the tasks are completed for three consecutive lists. Testing and guessing effects only emerged after six consecutive lists, suggesting that expectancy processes require considerable repetition before expectancies form, which likely alters how participants encode the list at study. The present experiments therefore supplement previous work demonstrating that testing and guessing tasks can improve memory performance, but that these improvements may be due at least in part to task-expectancy processes rather than the tasks themselves. Future research should examine how participant expectancies, particularly when participants have knowledge and practice of upcoming tasks or tests, may affect memory performance. Separating task effects from task-expectancy from retrieval-practice effects can provide researchers and educators alike with critical information about how and why task-related memory effects occur.

Note

1. Although we change to an online sample from an in-lab sample in Experiments 4A and 4B, we note that previous work employed both online and in-lab samples and showed similar effects using a similar paradigm (Huff et al., 2012, 2016). Indeed, Experiment 1 – which replicated Huff et al. (2016) but using an in-lab sample – similarly yielded no testing or guessing benefits over restudy. Therefore, we do not think a shift to an online sample would be a strong contributor to any differences found across experiments and if anything would enhance the generalizability of the findings given the mturk sample is more representative of the population.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

Funding support was provided by the National Institute on Aging (NIA; T32 AG0000-0-39).

References

Balota, D. A., & Neely, J. H. (1980). Test-expectancy and word-frequency effects in recall and recognition. *Journal of Experimental Psychology: Human Learning and Memory*, *6*, 576–587. doi:10.1037/0278-7393.6.5.576

Balota, D. A., Yap, M. J., Cortese, M. J., Hutchison, K. A., Kessler, B., Loftis, B., … Trieman, R. (2007). The English lexicon project. *Behavior Research Methods*, *39*, 445–459.

Bartlett, S. F. C. (1932/1967). *Remembering: A study in experimental and social psychology*. Cambridge: Cambridge University Press.

Battig, W. F., & Montague, W. E. (1969). Category norms of verbal items in 56 categories. A replication and extension of the Connecticut category norms . *Journal of Experimental Psychology*, *80*, 1–46. doi:10.1037/h0027577

Begg, I., & Snider, A. (1987). The generation effect: Evidence for generalized inhibition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *13*, 553–563. doi:10.1037/0278-7393.13.4.553

Bjork, R. A. (1975). Retrieval as a memory modifier: An interpretation of negative recency and related phenomena. In R. L. Solso (Ed.), *Information processing and cognition: The Loyola Symposium* (pp. 123–144). Hillsdale, NJ: Erlbaum.

Brainerd, C. J., & Reyna, V. F. (2002). Fuzzy-trace theory and false memory. *Current Directions in Psychological Science*, *11*, 164–169. doi:10.111/1467-8721.00192

Carpenter, S. K. (2011). Semantic information activated during retrieval contributes to later retention: Support for the mediator effectiveness hypothesis of the testing effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*, 1547–1552. doi:10.1037/a0024140

Coane, J. H., Huff, M. J., & Hutchison, K. A. (2016). The ironic effect of guessing: Increased false memory for mediated lists in younger and older adults. *Aging, Neuropsychology, and Cognition*, *23*, 283–303.

Coltheart, M. (1981). The MRC psycholinguistic database. *The Quarterly Journal of Experimental Psychology A: Human Experimental Psychology*, *33*, 497–505. doi:10.1080/14640748108400805

Congleton, A., & Rajaram, S. (2012). The origin of the interaction between learning method and delay in the testing effect: The roles of processing and conceptual retrieval organization. *Memory & Cognition*, *40*, 528–539. doi:10.3758/s13421-011-0168-y

Deese, J. (1959). On the prediction of occurrence of particular verbal intrusions in immediate recall. *Journal of Experimental Psychology*, *58*, 17–22. doi:10.1037/h0046671

Gallo, D. A. (2004). Using recall to reduce false recognition: Diagnostic and disqualifying monitoring. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*, 120–128. doi:10.1037/0278-7393.30.1.120

Gardiner, J. M., Java, R. I., & Richardson-Klavehn, A. (1996). How level of processing really influences awareness in recognition memory. *Canadian Journal of Experimental Psychology*, *50*, 114–122. doi:10.1037/1196-1961.50.1.114

Glanzer, M., & Adams, J. K. (1990). The mirror effect in recognition memory: Data and theory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *16*, 5–16. doi:10.1037/0278-7393.16.1.5

Grimaldi, P. J., & Karpicke, J. D. (2012). When and why do retrieval attempts enhance subsequent encoding? *Memory & Cognition*, *40*, 505–513. doi:10.3758/s13421-011-0174-0

Hays, M. J., Kornell, N., & Bjork, R. A. (2013). When and why a failed test potentiates the effectiveness of subsequent study. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*, 290–296. doi:10.1037/a0028468

Huelser, B. J., & Metcalfe, J. (2012). Making related errors facilitates learning, but learners do not know it. *Memory & Cognition*, *40*, 514–527. doi:10.3758/s13421-011-0167-z

Huff, M. J., Balota, D. A., & Hutchison, K. A. (2016). The costs and benefits of testing and guessing on recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *42*, 1559–1572. doi:10.1037/xlm0000269

Huff, M. J., & Bodner, G. E. (2013). When does memory monitoring succeed versus fail? Comparing item-specific and relational encoding in the DRM paradigm. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*, 1246–1256. doi:10.1037/a0031338

Huff, M. J., & Bodner, G. E. (2014). All varieties of encoding variability are not created equal: Separating variable processing from variable tasks. *Journal of Memory and Language*, *73*, 43–58. doi:10.1016/j.jml.2014.01.004

Huff, M. J., Coane, J. H., Hutchison, K. A., Grasser, E. B., & Blais, J. E. (2012). Interpolated task effects on direct and mediated false recognition: Effects of initial recall, recognition, and the ironic effect of guessing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*, 1246–1256. doi:10.1037/a0028476

Huff, M. J., Davis, S. D., & Meade, M. L. (2013). The effects of initial testing on false recall and false recognition in the social contagion of memory paradigm. *Memory & Cognition*, *41*, 820–831. doi:10.3758/s13421-013-0299-4

Huff, M. J., Meade, M. L., & Hutchison, K. A. (2011). Age-related differences in guessing on free and forced recall tests. *Memory (Hove, England)*, *19*, 317–330. doi:10.1080/09658211.2011.568494

Hunt, R. R., & Einstein, G. O. (1981). Relational and item-specific information in memory. *Journal of Verbal Learning and Verbal Behavior*, *20*, 497–514. doi:10.1016/S0022-5371(81)90138-9

Jacoby, L. L. (1991). A process dissociation framework: Separating automatic from intentional uses of memory. *Journal of Memory and Language*, *30*, 513–541. doi:10.1016/0749-596X(91)90025-F

Jacoby, L. L. (1994). Measuring recollection: Strategic versus automatic influences of associative context. In V. Carlo & M. Moscovitch (Eds.), *Attention and performance 15: Conscious and nonconscious information processing* (pp. 661–679). Cambridge, MA: The MIT Press.

Kang, S. H. K., Pashler, H., Cepeda, N. J., Rohrer, D., Carpenter, S. K., & Mozer, M. C. (2011). Does incorrect guessing impair fact learning? *Journal of Educational Psychology*, *103*, 48–59. doi: 10.1037/a0021977

Kay, H. (1955). Learning and retaining verbal material. *British Journal of Psychology*, *46*, 81–100. doi:10.1111/j.2044-8295.1955.tb00527.x

Kelley, C. M., & Sahakyan, L. (2003). Memory, monitoring, and control in the attainment of memory accuracy. *Journal of Memory and Language*, *48*, 704–721. doi:10/1016/S0749-596X(02)00504-1

Koriat, A., & Goldsmith, M. (1996). Monitoring and control processes in the strategic regulation of memory accuracy. *Psychological Review*, *103*, 490–517. doi:10.1037/0033-295X.103.3.490

Kornell, N., Hays, M. J., & Bjork, R. A. (2009). Unsuccessful retrieval attempts enhance subsequent learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*, 989–998. doi:10.1037/a0015729

Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, & Computers*, *28*, 203–208. doi:10.3758/BF03204766

Mandler, G. (1980). Recognizing: The judgment of previous occurrence. *Psychological Review*, *87*, 252–271. doi:10.1037/0033-295X.87.3.252

Mason, W., & Suri, S. (2012). Conducting behavioral research on Amazon's mechanical Turk. *Behavior Research Methods*, *44*, 1–23. doi:10.3758/s13428-011-0124-6

Masson, M. E. J. (2011). A tutorial on practiceal Bayesian alternative to null-hypothesis significance testing. *Behavioral Research Methods*, *43*, null–hypothesis.

McDermott, K. B. (1996). The persistence of false memories in list recall. *Journal of Memory and Language*, *35*, 212–230. doi:10.1006/jmla.196.0012

Meade, M. L., & RoedigerIII, H. L. (2006). The effect of forced recall on illusory recollection in younger and older adults. *The American Journal of Psychology*, *119*, 433–462. doi:10.2307/20445352

Morris, C. D., Bransford, J. D., & Franks, J. J. (1977). Levels of processing versus transfer appropriate processing. *Journal of Verbal Learning and Verbal Behavior*, *16*, 519–533. doi:10.1016/S0022-5371(77)80016-9

Neely, J. H., & Balota, D. A. (1981). Test-expectancy and semantic-organization effects in recall and recognition. *Memory & Cognition*, *9*, 283–300. doi:10.3758/BF03196962

Pansky, A., Goldsmith, M., Koriat, A., & Pearlman-Avnion, S. (2009). Memory accuracy in old age: Cognitive, metacognitive, and neurocognitive determinants. *European Journal of Cognitive Psychology*, *21*, 303–329. doi:10.1080/09541440802281183

Potts, R., & Shanks, D. R. (2014). The benefit of generating errors during learning. *Journal of Experimental Psychology: General*, *143*, 644–667. doi:10.1037/a0033194

Pu, X., & Tse, C.-S. (2014). The influence of intentional versus incidental retrieval practices on the role of recollection in test-enhanced learning. *Cognitive Processing*, *15*, 55–64. doi:10.1007/s10339-013-0580-2

Pyc, M. A., & Rawson, K. A. (2010). Why testing improves memory: Mediator effectiveness hypothesis. *Science*, *330*, 335. doi:10.1126/science.1191465

Rawson, K. A., & Dunlosky, J. (2011). Optimizing schedules of retrieval practice for durable and efficient learning: How much is enough? *Journal of Experimental Psychology: General*, *140*, 283–302. doi:10.1037/a0023956

RoedigerIII, H. L., & Karpicke D. J. (2006). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science*, *1*, 181–210. doi:10.1111/j.1745-6916.2006.00012.x

RoedigerIII, H. L., & McDermott B. K. (1995). Creating false memories: Remembering words not presented in lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*, 803–814. doi:10.1037.0278-7393.21.4.803

RoedigerIII, H. L., & Payne, G. D. (1985). Recall criterion does not affect recall level or hypermnesia: A puzzle for generate/recognize theories. *Memory & Cognition*, *13*, 1–7. doi:10.3758/BF03198437

RoedigerIII, H. L., Wheeler, M. A., & Rajaram, S. (1993). Remembering, knowing and reconstructing the past. In D. L. Medin (Ed.), *The psychology of learning and motivation* (pp. 97–134). San Diego, CA: Academic Press.

Rowland, C. A. (2014). The effects of testing versus restudy on retention: A meta-analytic review of the testing effect. *Psychological Bulletin*, *140*, 1432–1463. doi:10.1037/a0037559

Slamecka, J. J., & Graf, P. (1978). The generation effect: Delineation of a phenomenon. *Journal of Experimental Psychology: Human Learning and Memory*, *4*, 592–604. doi:10.1037/0278-7393.4.6.592

Szpunar, K. K., McDermott, K. B., & RoedigerIII, H. L. (2007). Expectation of a final cumulative test enhances long-term retention. *Memory & Cognition*, *35*, 1007–1013. doi:10.3758/BF03193473

Tse, C.-S., Balota, D. A., & RoedigerIII, H. L. (2010). The benefits and costs of repeated testing on the learning of face-name pairs in healthy older adults. *Psychology and Aging*, *25*, 833–845. doi:10.1037/a0019933

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*, 289–335. doi:10.1016/j.jml.2003.10.003

Wagenmakers, E.-J. (2007). A practical solution to the pervaise problems of p values. *Behavioral Research Methods*, *43*, 679–690. doi:10.3758/s13428-10-0049-5

Yan, V. X., Yu, Y., Garcia, M. A., & Bjork, R. A. (2014). Why does guessing incorrectly enhance, rather than impair, retention? *Memory & Cognition*, *42*, 1373–1383. doi:10.3758/s13421-014-0454-6

Yonelinas, A. P. (2002). The nature of recollection and familiarity: A review of 30 years of research. *Journal of Memory and Language*, *46*, 441–517. doi:10.1006/jmla.2002.2864

Zaromb, F. M., & RoedigerIII, H. L. (2010). The testing effect in free recall is associated with enhanced organizational processes. *Memory & Cognition*, *38*, 995–1008. doi:10.3758/MC.38.8.995