Journal of Experimental Psychology: Learning, Memory, and Cognition

The Costs and Benefits of Testing and Guessing on Recognition Memory

Mark J. Huff, David A. Balota, and Keith A. Hutchison

Online First Publication, March 7, 2016. http://dx.doi.org/10.1037/xlm0000269

Journal of Experimental Psychology:
Learning, Memory, and Cognition

0278-7393/16/$12.00 http://dx.doi.org/10.1037/xlm0000269

The Costs and Benefits of Testing and Guessing on Recognition Memory

Mark J. Huff and David A. Balota
Washington University in St. Louis

Keith A. Hutchison
Montana State University

We examined whether 2 types of interpolated tasks (i.e., retrieval-practice via free recall or guessing a missing critical item) improved final recognition for related and unrelated word lists relative to restudying or completing a filler task. Both retrieval-practice and guessing tasks improved correct recognition relative to restudy and filler tasks, particularly when study lists were semantically related. However, both retrieval practice and guessing also generally inflated false recognition for the nonpresented critical words. These patterns were found when final recognition was completed during a short delay within the same experimental session (Experiment 1) and after a 24-hr delay (Experiment 2). In Experiment 3, task instructions were presented randomly after each list to determine whether retrieval-practice and guessing effects were influenced by task-expectancy processes. In contrast to Experiments 1 and 2, final recognition after retrieval practice and guessing was equivalent to restudy, suggesting that the observed retrieval-practice and guessing advantages were in part because of preparatory task-based processing during study.

Keywords: guessing, retrieval-practice, recognition, free recall, retention interval

Memory researchers have long sought methods to enhance retention and many successful techniques often target how individuals process to-be-remembered information during encoding. Such techniques include elaborative or "deep" processing (Craik & Lockhart, 1972; see Craik, 2002, for a review), item generation (Slamecka & Graf, 1978), and more recently, processing items for survival advantages (Nairne, Thompson, & Pandeirada, 2007). Researchers have also become increasingly interested in how varieties of interpolated tasks—tasks that vary the degree with which the learner engages with the material after study has occurred—can also improve retention on a subsequent memory task. A highly replicable and robust example of an interpolated-task benefit is referred to as the testing or retrieval-practice effect (see Rawson & Dunlosky, 2011, for a review; Roediger & Karpicke, 2006), in which memory is better when a test is completed after study relative to a restudy control group. These interpolated-task benefits demonstrate that memory can be enhanced when individuals actively engage with the material after initial encoding. The purpose of our study is to investigate how another type of interpolated task, attempting to guess words related to those previously studied, may also act as a method to improve retention similar to that of retrieval practice.

Mark J. Huff and David A. Balota, Department of Psychology, Washington University in St. Louis; Keith A. Hutchison, Department of Psychology, Montana State University.

Funding support was provided by the National Institute on Aging (NIA; T32 AG0000-0-39). We express appreciation to Katie Hebeler for assistance with data collection and coding.

Correspondence concerning this article should be addressed to Mark J. Huff, Department of Psychology, Washington University in St. Louis, St. Louis, MO 63130. E-mail: mhuff@wustl.edu

Guessing in the current context refers to identifying and reporting specific items one knows were not studied. This differs from a common usage of the term guessing in memory experiments, which refers instead to reporting an item as studied based on little-to-no memory for its occurrence. This latter form of guessing is likely a common test strategy used to increase the number of correct items retrieved, though often results in a cost when errors are factored into calculating overall memory accuracy (Huff, Meade, & Hutchison, 2011; Meade & Roediger, 2006). Certain conditions encourage guessing, such as when participants are told there are no penalties for guessing (Kelley & Sahakyan, 2003; Koriat & Goldsmith, 1996), under conditions of forced report (Meade & Roediger, 2006; Roediger & Payne, 1985), and when participants are given categorized (vs. unrelated) test materials (Huff et al., 2011; Meade & Roediger, 2006, 2009). In addition, there is some evidence that adopting a liberal response criterion, which likely increases guessing, may operate as a stable cognitive trait that is consistent across study-test cycles and across different study materials (Bengson & Hutchison, 2007; Kantner & Lindsay, 2012). Thus, guessing occurs frequently over a broad range of test materials and test instructions, and when guessing does occur, it does not always increase overall memory accuracy (see Roediger, Wheeler, & Rajaram, 1993 for a review).

Given that participants often guess based on partial information, it is important to consider how guessing affects subsequent retrieval in paradigms that require the completion of repeated tests such as in retrieval-practice studies. This is particularly important if guessing on initial tests results in memory errors on subsequent tests. For instance, previous research has shown that, in the absence of corrective feedback, errors made on initial tests often persist on later tests (Bartlett, 1932/1967; Huff, Davis, & Meade, 2013; Kang, Pashler, Cepeda, Rohrer, Carpenter, & Mozer, 2011; Kay, 1955; Lane, Mather, Villa, & Morita, 2001; McDermott, 1996; Roediger, Jacoby, & McDermott, 1996; Tse, Balota, &

Roediger, 2010). Therefore, a natural concern is how such intrusions may affect memory benefits characteristic of repeated-test paradigms. These errors may be particularly problematic given the memory-enhancing processes of generation (i.e., Slamecka & Graf, 1978) as an intrusion yields a self-produced error. Perhaps this possibility is why learning theorists have long supported the idea of errorless learning in which learning is greatest under conditions that minimize or eliminate errors (e.g., Baddeley & Wilson, 1994; Skinner, 1953, 1958).

In contrast to the credo of errorless learning, researchers have shown that guessing may sometimes be advantageous to later memory when corrective feedback is provided. For example, Kornell, Hays, and Bjork (2009; see too Grimaldi & Karpicke, 2012; Huelser & Metcalfe, 2012) presented participants with a single cue word with instruction to guess a related target word. For all pairs, the cue-target relation was weak, which made successful production of the target word difficult and largely unsuccessful (only approximately 4–5% of guesses were correct targets). After participants provided a target response, they were immediately presented with corrective feedback by viewing a fully intact cue-target pair. These guess trials were compared with interleaved read-only trials in which intact cue-target pairs were simply studied. A final cued-recall test was then completed in which the original cue word was presented with instruction to retrieve the previously studied target. Critically, final test performance was higher when participants attempted to guess the target word than when reading an intact pair. This pattern was found when the final test occurred 5 min after study (Experiment 4), after a 24-hr delay (Experiment 5; see too Yan, Yu, Garcia, & Bjork, 2014), and when guess versus intact pairs were presented between-subjects (Experiment 6).

One account of guessing benefits is that they occur because of retrieval-based elaborative processes in which generating an incorrect target in response to a cue (that is likely semantically related) may function as a mediator that would enhance associations to other concepts that could be used as retrieval cues (Grimaldi & Karpicke, 2012; Hays, Kornell, & Bjork, 2013; Huelser & Metcalfe, 2012; Knight, Ball, Brewer, DeWitt, & Marsh, 2012; Kornell et al., 2009). These cues could be used as reminders of the intact cue-target pair provided as feedback or, if only the incorrect guess could be retrieved, act as a mediator cue for the intact pair stored in memory (i.e., mediator effectiveness hypothesis; Pyc & Rawson, 2010). In either case, with semantically related pairs, incorrect guessing can enrich the associative characteristics between the cue and target when participants are presented with the intact pair during feedback to promote retrieval. While this possibility nicely accounts for guessing benefits when study materials share pre-experimental semantic associations, it may not apply to guessing benefits that occur when materials are unrelated. Indeed, there is some evidence that even with word pairs that do not possess pre-experimental associations, such as foreign language pairs, the benefits of guessing remain with feedback (e.g., Potts & Shanks, 2014) in contrast to the mediator account.

Although guessing studies have reliably shown robust memory benefits for targets in cued-recall studies when feedback is given, it is unclear whether guessing may also enhance memory for the information that is used to generate the target word initially (i.e., the cue word). In previous guessing studies (e.g., Kornell et al., 2009; Potts & Shanks, 2014), the cue word is always presented at test which precludes an assessment of whether guessing enhances memory for the cue word itself. Given that guessing likely either increases semantic mediators between cue-target pairs or increases attentional processing given to corrective feedback, it is likely that guessing also improves memory for the initially presented items.

Relevant to our study, the effects of guessing have also been examined in associative-list paradigms in which participants are asked to guess a nonpresented critical lure after studying a list of associates. Specifically, Huff, Coane, Hutchison, Grasser, and Blais (2012; see too Huff & Hutchison, 2011; Coane, Huff, & Hutchison, in press) presented participants with lists of directly related words (e.g., *water*, *bridge*, *run*, etc.) or mediated lists consisting of indirectly related words (e.g., *faucet*, *London*, *jog*, etc.) that converged upon a nonpresented critical lure (e.g., *river*). Directly related lists consisted of associates used in the popular Deese-Roediger-McDermott (DRM; Deese, 1959; Roediger & McDermott, 1995) false memory paradigm. Mediated lists consisted of the indirectly related associates for which mediators converged upon the critical lure. Participants first studied a list that was immediately followed by a between-subjects interpolated task consisting of either an arithmetic filler task, a free-recall test, or a guessing task in which participants were asked to guess the nonpresented critical lure. After several study-filler/recall/guess cycles, participants completed a final recognition test. Replicating past research, correct recognition on the final test was greater after interpolated recall than interpolated arithmetic. Importantly, however, it was greatest after the interpolated guessing for both DRM and mediated lists even though participants were highly inaccurate in guessing critical lures from mediated lists. This pattern is important in light of previous guessing experiments (e.g., Kornell et al., 2009; Potts & Shanks, 2014) because Huff et al.'s guess group did not receive feedback in the form of the representation of study items. In other words, attempting to generate a critical item that was not presented produced a final recognition benefit greater than retrieval practice of list items that were presented.

In addition to memory for the list items, Huff et al. found that false recognition of critical lures was also moderated by the type of interpolated task completed. For DRM lists, false recall was greatest after interpolated recall and lower and equal after arithmetic and guessing tasks. Guessing could presumably act as an implied warning, because participants successfully guessed critical lures 40% of the time and could therefore later monitor for those lures on the final recognition test. For mediated lists, however, critical lures were only successfully guessed on 5% of lists. For these lists, false recognition of critical items was greater after guessing than either recall or arithmetic groups—a pattern termed an *ironic effect of guessing*. Therefore, interpolated guessing without feedback appears to improve correct recognition and can either decrease or increase critical lure false recognition, depending on how well the critical lure can be identified.

In the present study, we sought to replicate and extend the effects of guessing reported by Huff et al. (2012) by more closely examining the influence of interpolated tasks on final correct recognition. Guessing may operate as a type of internal retrieval, as participants mentally retrieve and associate the study items to generate the critical lure. This type of internal retrieval may be as (or even more) effective than completing an explicit memory test that requires participants to report studied items. To be more consistent with work examining retrieval practice, we included a

restudy interpolated group in which participants were re-exposed to the items that were originally studied. If attempting to guess items that were not originally presented on a study list enhances final recognition, it is important to determine whether this benefit persists over and above restudy. Therefore, an appropriate control is a group that is represented with the study list (vs. a filler task). Hence, we compared filler task, restudy, retrieval-practice (i.e., free recall), and guess interpolated tasks and their subsequent effects on final correct and false recognition. The filler, retrieval-practice, and guess tasks closely matched those used by Huff et al. and the restudy group was given a second study opportunity of the original list in a different order.

To gain greater traction on the types of materials that may elicit guessing benefits, we compared lists that were either strongly related through category membership (e.g., fruits, animals, etc.), weakly related through broad categories (i.e., ad hoc lists; e.g., things made of wood, things that are green, etc.; Huff & Bodner, 2014; Hunt & Einstein, 1981), and lists that were unrelated. In Huff et al. (2012), interpolated guessing benefits were found on both directly related DRM lists and on weakly related mediated lists; however, it is unclear whether these benefits are restricted to lists that share some form of association, or if the benefits are more general and can be found with unrelated and weakly related ad hoc lists. The benefits of guessing with cue-target pairs may be limited in some cases to the association between the items (see Potts & Shanks, 2014, for discussion) and, therefore, it is important to determine whether list association operates as a boundary condition in a list-learning paradigm.

In Experiment 1, participants were presented with categorized, ad hoc, or unrelated lists. Immediately after study, participants completed one of four interpolated tasks that consisted of either a filler task, a restudy opportunity, retrieval practice through a free-recall test, or a guessing task in which they were asked to guess up to five nonpresented critical items that were related to the list on some dimension. Of note, interpolated task instructions were presented before participants began studying the lists and were a between-subjects manipulation. After study, all groups were given 60 s to complete the interpolated task with the exception of the guess group, which was given up to 60 s to guess the critical items. Groups completed six study/interpolated task cycles after which they received a final recognition test.

If the guessing benefit observed by Huff et al. (2012) extends to these new materials, one should find that the interpolated guess group would produce the highest level of correct recognition followed by the retrieval practice group, the restudy group, and finally the filler group. In addition, based on previous results using paired associates and corrective feedback, one may expect a guessing benefit will be found with categorized lists (cf. Kornell et al., 2009; Yan et al., 2014), ad hoc lists (cf. Grimaldi & Karpicke, 2012; Huelser & Metcalfe, 2012; Yan et al., 2014), and unrelated lists (cf. Potts & Shanks, 2014).

For false recognition, guessing was expected to either decrease or increase false recognition depending on the relatedness of the study list. For categorized lists, guessing was expected to reduce false recognition relative to retrieval practice, to the same level as the filler group. This is because guessing acts as an implied warning that related lures exist and also, provided guessing is successful, allows participants to use a recall-to-reject strategy during later recognition (Gallo, 2004). Restudy is expected to produce the lowest level of false recognition, as repeated study of DRM lists has been shown to reduce false recognition (Benjamin, 2001; Seamon, Luo, Schwartz, Jones, Lee, & Jones, 2002). For ad hoc critical items, guessing is expected to produce the highest level of false recognition followed by the retrieval-practice and filler groups—consistent with the ironic effect of guessing using mediated lists (Huff et al., 2012). Once again, restudy is expected to produce the lowest level of false recognition because of repeated study effects in associative false memory. Critical item false recognition on unrelated lists is not expected to show differences across interpolated tasks given these items are essentially unrelated control items and would not be affected by processes recruited by the interpolated tasks.

Experiment 1: Immediate Recognition

Method

Participants. One hundred fifty-six individuals were recruited using Amazon's Mechanical Turk (for an overview, see Mason & Suri, 2012) and randomly assigned to the filler ($N = 39$), restudy ($N = 40$), retrieval-practice ($N = 39$), and guessing ($N = 38$) groups. All reported proficiency in English, resided in the United States, and had normal or corrected-to-normal vision. Mean reported age was 34.81 years ($SD = 11.98$, range = 18–65) and mean reported formal education was 15.63 years ($SD = 2.16$, range = 10–24). Two participants reported cheating during the experiment and were replaced.

Materials. Two sets of categorized, ad hoc, and unrelated word lists were constructed. For categorized lists, each set contained items from four Battig and Montague (1969) categories (Set A = four-footed animals, tools, fruits, and spices; Set B = birds, furniture, vegetables, and human body parts). The top 25 typical exemplars from each category were used. Of these 25 exemplars, the top 5 (i.e., the most typical exemplars) were designated as critical items and, therefore, were not studied and the remaining 20 were used as study items for each list. For ad hoc lists, two sets were similarly created such that each contained four broad categories (Set A = things that are green, things that make noise, liquids, and things that are soft; Set B = things that are black, things made of wood, things in a kitchen, and things women wear; Hunt & Einstein, 1981; Van Overschelde, Rawson, & Dunlosky, 2004). As in the categorized lists, the top 25 exemplars from each category were taken and the top 5 were designated as critical items and the remaining 20 as study items (see Huff & Bodner, 2014 for a similar procedure). Categorized list items were found to be both longer in word 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), $ts > 2.31$, $ps < .01$. Categorized and ad hoc lists were similar in concreteness and familiarity in the MRC Psycholinguistic Database (Coltheart, 1981), $ts < 1.70$, $ps > .10$. Finally, 200 randomly selected unrelated words were generated, which were taken from the MRC database and were matched in word length, frequency, concreteness, familiarity, and length to the ad hoc lists, $ts > 1.96$, $ps > .05$. Unrelated items were not listed as members of the categories used in categorized lists or as members from the broad ad hoc categories. Unrelated items were divided into 8 separate lists of 25 items that were used to make up two sets of four lists.

Within each list, 5 were randomly designated as nonpresented critical items and the remaining 20 were studied list items. Thus, each set contained 12 total lists consisting of 4 lists from each of the categorized, ad hoc, and unrelated list types. Presentation of study list items was randomized anew for each participant.

Each 12-list set was further subdivided into two study-test blocks that consisted of six presented study lists (two from each list type) that were presented in the same preset order for each block (categorized, ad hoc, unrelated, categorized, ad hoc, unrelated). A 180-item recognition test was presented in a newly randomized order for each participant at the end of each study block. Each test was composed of 60 studied list items (10 from each of the 6 lists from even-numbered serial positions, making 20 categorized, ad hoc, and unrelated items), 60 nonpresented list items from the other nonpresented list set (from the same positions and list types), 30 critical items from studied lists (5 from each of the 6 lists, making 10 categorized, ad hoc, and unrelated items), and 30 critical items from nonpresented lists (5 per nonpresented list from the same list types). A second 180-item recognition test was created using the same format for the second block of lists. The order of the two study-test blocks was counterbalanced across participants.

Procedure. Participants were provided with a link that directed them to an Internet-based program that provided an electronic consent form, study and test materials, and a brief demographics form. After providing consent, all participants received written instructions that they would study two blocks of 6 word lists and that each list contained 20 words that would be individually displayed for 3 s. Thus, each list was presented within the span of 60 s. Participants were further informed that their memory for these lists would later be tested and asked not to write down any of the words as they were presented and informed that compensation for the experiment was not tied to performance. Participants were not provided with any information about the type of memory test they would complete.

Following the general instructions, participants were randomly assigned to one of four interpolated-task groups and received task-relevant instructions before the initial study phase. Words were presented in a large 48-pt sans serif font in the center of the screen in all caps. In the filler group, participants were instructed that they would complete a Tetris filler task for 60 s after the presentation of the list. In the restudy group, participants were instructed that they would be presented with the same word list they just studied (3 s per word; 60 s total) but in a different randomized order. In the retrieval-practice group, participants were instructed that immediately after study of each list they would be asked to freely recall as many items from the list they studied by typing their responses into a dialog box on the screen for 60 s without cost for misspellings. In the guess group, participants were told that for each list, five words that were associated to the list on some dimension were not presented. Following each list, their task was to attempt to guess up to five of these nonpresented words by typing them into a dialog box on the screen. Guess participants were also informed that they would have up to 60 s to try to guess the words but could advance to the next list if they could not think of other nonpresented words by clicking a submit button below the dialog box. Participants were required to provide at least one guess for each list. The time between the appearance of the guessing dialog box to when participants clicked the submit button was measured to determine the amount of time spent entering guesses.

Immediately after the sixth list in the first block, an old/new recognition test was completed. A word was presented on the center of the computer screen (in the same font/size used at study) along with two buttons referring to old and new. Participants were instructed to click the old button with the mouse if the word was studied on the previous six lists and the new button if the word was not. The word remained on the screen until participants made a response. The second block of lists began immediately after the recognition test, with a reminder of the interpolated instructions completed in the first block. The second study block was followed by a second recognition test, a demographics questionnaire, and debriefing information. The demographic questionnaire asked participants to report their age, years of formal education, whether they completed the task with normal or corrected vision, and if they cheated during the task. For the cheating question, participants were directly asked if they cheated during the experiment such as writing down the study list as the words were being displayed. They were asked to answer the question honestly and told that even if they reported cheating they would still be compensated. The experiment lasted approximately 45 min and participants were compensated $2.00 for completion awarded through amazon.com.

Results

A $p < .05$ level of significance was used unless otherwise noted. Effect size measures using η_p^2 and Cohen's d are reported for all significant effects from the analyses of variances (ANOVAs) and t tests, respectively. For all nonsignificant comparisons reported, we further tested these comparisons using a Bayesian estimate of the strength of evidence supporting the null hypothesis (Masson, 2011; Wagenmakers, 2007). In this analysis, two models are compared. One model assumes an effect and is compared with a model that assumes a null effect. This Bayesian analysis yields a probability estimate that the null difference is retained—a p value termed p_{BIC} (Bayesian Information Criterion). This calculation is sensitive to the sample size and can act as a power analysis to increase confidence in the reported null effect. Thus, we utilize the p_{BIC} analysis to supplement null effects found using traditional null-hypothesis-significance testing.

Interpolated recall performance. Interpolated-task performance for the retrieval-practice and guessing groups is reported in Table 1. We performed separate one-way ANOVAs on the proportion of correctly recalled list items and falsely recalled critical item intrusions for participants in the recall group. Correct recall differed across list types, $F(2, 76) = 44.18$, $MSE = .01$, $\eta_p^2 = .54$. Relative to categorized lists, recall was lower in both the ad hoc (.57 vs. .52), $t(38) = 4.13$, $SEM = .02$, $d = 0.39$, and unrelated lists (.57 vs. .41), $t(38) = 7.50$, $SEM = .02$, $d = 1.17$. Correct recall was also greater in the ad hoc than unrelated lists (.52 vs. .41), $t(38) = 6.30$, $SEM = .02$, $d = 0.78$. Therefore, as expected, lists that provided the strongest semantic structure also produced the greatest correct recall.

False recall was also found to differ across list types, $F(2, 76) = 12.62$, $MSE = .01$, $\eta_p^2 = .25$. Critical items were recalled more frequently on categorized than ad hoc lists (.03 vs. .01), $t(38) = 3.06$, $SEM = .01$, $d = 0.65$, and on categorized than unrelated lists

Table 1
Mean (SD) Initial Recall Proportions for List Items and Critical Items and Correct Guessing of Critical Items of Categorized, Ad Hoc, and Unrelated Lists in Experiments 1–3

	Interpolated task group	
Experiment/item type	Recall	Guess
Experiment 1		
List items		
Categorized	.57 (.12)	—
Ad hoc	.52 (.14)	—
Unrelated	.41 (.15)	—
Critical items		
Categorized	.03 (.05)	.20 (.15)
Ad hoc	.01 (.02)	.11 (.13)
Unrelated	.00 (.00)	.00 (.00)
Experiment 2		
List items		
Categorized	.58 (.10)	—
Ad hoc	.54 (.11)	—
Unrelated	.47 (.14)	—
Critical items		
Categorized	.02 (.02)	.21 (.16)
Ad hoc	.02 (.04)	.13 (.12)
Unrelated	.00 (.00)	.00 (.00)
Experiment 3		
List items		
Categorized	.43 (.15)	—
Ad hoc	.42 (.21)	—
Unrelated	.29 (.22)	—
Critical items		
Categorized	.06 (.12)	.23 (.26)
Ad hoc	.01 (.05)	.05 (.15)
Unrelated	.00 (.00)	.00 (.00)

(.03 vs. .00), $t(38) = 4.11$, $SEM = .01$, $d = 0.93$. False recall was also greater on ad hoc than unrelated lists (.01 vs. .00), $t(38) = 2.63$, $SEM = .01$, $d = 0.60$, though false recall was very low across all list types.

Interpolated guessing performance. The proportion of correctly guessed critical items for participants in the initial guessing condition were calculated by taking the total number of critical items correctly guessed divided by the total number possible and were analyzed similarly to correct recall. Correctly guessed items were found to differ across list types, $F(2, 74) = 39.24$, $MSE = .01$, $\eta_p^2 = .52$. Critical items were successfully guessed more frequently on categorized than ad hoc lists (.20 vs. .11), $t(37)$ 3.97, $SEM = .02$, $d = 0.62$, and on categorized than unrelated lists (.20 vs. .00), $t(37) = 8.13$, $SEM = .02$, $d = 1.86$. Correct guessing was also more frequent on ad hoc than unrelated lists (.11 vs. .00), $t(37) = 5.35$, $SEM = .02$, $d = 1.23$. Thus, again as expected, successful guessing was also aided when study list items shared semantic relations.

We also tabulated the amount of time the guessing dialog box was available on the computer screen for each list to gauge the amount of time participants spent attempting to guess the critical items. On average, participants spent 18.46 s per list guessing critical items, which is less than a third of the time granted to the retrieval-practice and restudy groups to complete their interpolated tasks.

Recognition. Proportions of studied list items and nonstudied critical items from categorized, ad hoc, and unrelated lists that were given an "old" response are reported in Table 2. Means are separated for each of the interpolated tasks (filler, restudy, retrieval-practice, and guess). Recognition proportions were adjusted using a hits minus false alarms correction for correct recognition (hit rates to studied list items minus false alarms to nonpresented controls), and critical item false recognition (false alarm rates to critical items from studied lists minus control critical items from nonpresented lists).[1]

A 3 (List Type) × 4 (Interpolated Task) mixed factorial ANOVA was used to contrast the effects of list and task types on correct recognition of list items. As was found in correct recall, correct recognition differed across list types, $F(2, 304) = 105.89$, $MSE = .01$, $\eta_p^2 = .41$, with recognition greater on categorized than ad hoc lists (.69 vs. .62), $t(155) = 6.96$, $SEM = .02$, $d = 0.33$, and on categorized than unrelated lists (.69 vs. .53), $t(155) = 12.99$, $SEM = .02$, $d = 0.75$. Correct recognition was also greater on ad hoc than unrelated lists (.62 vs. .53), $t(155) = 8.24$, $SEM = .02$, $d = 0.44$.

More important, correct recognition was also influenced by the interpolated task completed, $F(3, 152) = 6.87$, $MSE = .11$, $\eta_p^2 = .12$. Consistent with retrieval-practice benefits, recognition was greater in the retrieval-practice group than both the restudy (.70 vs. .57), $t(77) = 2.93$, $SEM = .03$, $d = 0.67$, and filler groups (.70 vs. .53), $t(76) = 4.00$, $SEM = .03$, $d = 0.92$. Critically, these same benefits were also found in the guessing group where guessing increased recognition relative to the restudy group (.66 vs. .57), $t(76) = 2.22$, $SEM = .03$, $d = 0.51$, and the filler group (.66 vs. .53), $t(76) = 3.29$, $SEM = .03$, $d = 0.76$. Interpolated guessing was not advantageous over retrieval-practice however (.66 vs. .70), $t(75) = 1.05$, $SEM = .02$, $p = .30$, $p_{BIC} = .83$, demonstrating that attempting to guess items that were not presented was similarly effective at enhancing recognition as recalling items that were presented. Recognition in the restudy group did not differ from the filler group (.57 vs. .53), $t < 1$, $p_{BIC} = .87$. The Interpolated Task × List Type interaction was not significant, $F(2, 304) = 1.79$, $MSE = .01$, $p = .10$, $p_{BIC} = .99$, demonstrating that the guessing benefit over restudy was general; occurring across materials that possessed a variety of associations.

Turning to critical item false recognition, proportions were analyzed as in correct recognition. False recognition differed across list types, $F(2, 304) = 140.18$, $MSE = .02$, $\eta_p^2 = .48$. As in false recall, false recognition was greater on categorized than ad hoc lists (.24 vs. .15), $t(155) = 6.78$, $SEM = .01$, $d = 0.53$, and categorized than unrelated lists (.24 vs. −.01), $t(155) = 14.02$, $SEM = .02$, $d = 1.61$. False recognition was also greater on ad hoc than unrelated lists (.15 vs. −.01), $t(155) = 10.88$, $SEM = .01$, $d = 1.24$. False recognition also differed across interpolated task groups, $F(3, 153) = 3.00$, $MSE = .03$, $\eta_p^2 = .06$. Consistent with the ironic effect of guessing, false recognition was inflated in the guess group relative to the filler group (.17 vs. .10), $t(75) = 2.89$, $SEM = .02$, $d = 0.68$, and relative to the restudy group (.17 vs. .11), $t(76) = 2.27$, $SEM = .02$, $d = 0.52$, but not relative to the

[1] A signal-detection analysis that computed *d*' for list items and critical items (Snodgrass & Corwin, 1988; Wickens, 2002) was also conducted in addition to the corrected proportions reported. All statistical patterns for list item and critical item corrected recognition in Experiments 1–3 were identical to those in the *d*' analysis. Analyses on corrected recognition are only reported to eliminate redundancy.

Table 2
Mean (SD) Recognition Proportions for List Items and Critical Items of Categorized, Ad Hoc, and Unrelated Lists as a Function of Initial Testing Conditions for Experiment 1

Item type	Interpolated task group			
	Filler	Restudy	Recall	Guess
N	39	40	39	38
List items				
Categorized	.79 (.12)	.84 (.12)	.89 (.10)	.88 (.08)
Controls	.18 (.19)	.21 (.20)	.08 (.10)	.14 (.12)
Ad hoc	.74 (.15)	.78 (.16)	.80 (.13)	.82 (.13)
Controls	.18 (.16)	.21 (.16)	.11 (.09)	.15 (.11)
Unrelated	.59 (.18)	.71 (.18)	.71 (.13)	.73 (.17)
Controls	.17 (.16)	.20 (.19)	.11 (.10)	.15 (.11)
Corrected recognition				
Categorized	**.61 (.24)**	**.63 (.25)**	**.80 (.16)**	**.74 (.15)**
Ad hoc	**.56 (.22)**	**.57 (.22)**	**.70 (.19)**	**.67 (.15)**
Unrelated	**.43 (.24)**	**.51 (.27)**	**.60 (.17)**	**.58 (.16)**
Task average	**.53 (.21)**	**.57 (.23)**	**.70 (.15)**	**.66 (.13)**
Critical items				
Categorized	.38 (.22)	.45 (.27)	.39 (.22)	.45 (.20)
Controls	.20 (.18)	.24 (.22)	.12 (.12)	.15 (.11)
Ad hoc	.27 (.18)	.33 (.23)	.21 (.17)	.32 (.16)
Controls	.15 (.18)	.20 (.19)	.09 (.11)	.14 (.11)
Unrelated	.16 (.14)	.16 (.18)	.09 (.14)	.12 (.13)
Controls	.16 (.16)	.17 (.18)	.10 (.11)	.11 (.11)
Corrected recognition				
Categorized	**.18 (.19)**	**.21 (.19)**	**.27 (.20)**	**.31 (.20)**
Ad hoc	**.12 (.13)**	**.14 (.16)**	**.13 (.15)**	**.21 (.16)**
Unrelated	**.00 (.08)**	**−.01 (.11)**	**−.01 (.11)**	**−.02 (.10)**
Task average	**.10 (.09)**	**.11 (.11)**	**.13 (.11)**	**.17 (.11)**

Note. Boldface indicates data used in the statistical analyses.

retrieval-practice group (.17 vs. .13), $t(75) = 1.37$, $SEM = .02$, $p = .17$, $p_{BIC} = .77$. No other group comparisons were significant, $ts > 1.41$, $ps > .16$, $p_{BIC}s > .77$. Unlike correct recognition, an Interpolated Task × List Type interaction was found, $F(6, 304) = 2.96$, $MSE = .02$, $\eta_p^2 = .06$, which reflected relatively low rates of false recognition in the filler group that increased across the restudy and retrieval-practice groups and greatest in the guess group, but only for the categorized and ad hoc lists (an ironic effect of guessing, Huff et al., 2012). For unrelated lists, however, false recognition did not change across interpolated task groups. Thus, when study lists were related, guessing produced the highest rate of false recognition.

Discussion

The results from Experiment 1 are quite clear. The type of interpolated task completed after the list presentation had a strong effect on final recognition. Specifically, for correct recognition of list items, interpolated retrieval-practice and guessing tasks increased correct recognition greater than both the restudy and filler groups. This benefit occurred across categorized, ad hoc, and unrelated list types demonstrating a general benefit across materials. It is also interesting that participants spent considerably less time completing the interpolated guessing task than the retrieval-practice group, which may suggest that guessing is an efficient method for improving later recognition. Of course, it is unclear whether retrieval-practice participants spent the entire 60 s actively retrieving list items; however, it is likely that they spent more time than the 18 s that participants spent guessing potential critical items.

Turning to the nonpresented critical items, interpolated false recall was quite low, though as expected, was greatest for categorized lists followed by ad hoc then unrelated lists—a pattern echoed in correct guessing of critical items. On the recognition test, categorized lists also produced the greatest level of false recognition followed by ad hoc and unrelated lists. We further showed that interpolated task types strongly affected false recognition for critical items. Consistent with the ironic effect of guessing, false recognition was greatest in the guess group followed by the retrieval-practice group, and lowest in the restudy and filler groups. This pattern was expected for ad hoc lists, given their relatively weak relations within the broad category membership akin to mediated lists, but unexpected for categorized lists as participants were more successful at guessing categorized critical items that could then be monitored for on the final recognition test. One reason for this unexpected finding is that, although participants were relatively more successful at correctly guessing critical items from categorized lists, successful guessing was still relatively low (20%) and may need to be sufficiently high before a reduction in false recognition is found. In the case of DRM lists, guessing is more likely given all list items converge upon a single lure whereas categorized list items do not. In summary, it appears that although guessing is beneficial to correct recognition relative to a restudy control task, it also produces a cost by inflating false recognition to related critical items.

Experiment 2: Delayed Recognition

In Experiment 2, we further compared guess, retrieval-practice, restudy, and filler interpolated tasks on correct recognition by evaluating task effects over a 24 to 48 hr retention interval, a common manipulation in retrieval-practice studies. Retrieval-practice benefits are generally found to be greater after a delay (vs. no delay), typically because of a greater increase in forgetting in the restudy-control group relative to the retrieval-practice group (Rawson & Dunlosky, 2011; Roediger & Karpicke, 2006). It is possible that guessing benefits relative to restudy may only occur when the final test is completed over a short retention interval in the same experimental session as study. One possibility is that guessing may be more susceptible to forgetting than retrieval practice, which may occur because guessing does not require participants to explicitly retrieve studied list items. Therefore, Experiment 2 inserted a long delay between the final study/interpolated task cycle and the final recognition test to evaluate whether the benefits of interpolated guessing persist after a longer retention interval.

Method

Participants. Seventy-seven individuals were recruited using Amazon's Mechanical Turk and randomly assigned to the filler ($N = 20$), restudy ($N = 19$), retrieval-practice ($N = 18$), and guess ($N = 20$) groups. All were proficient in English, resided in the United States, and had normal or corrected-to-normal vision. Because of a programming error, age and education information was not recorded for participants. No participants reported cheating during the experiment.

Materials and procedure. All procedures and stimulus lists were identical to those of Experiment 1 with two exceptions. First,

participants studied all 12 lists and the assigned interpolated task in a single block, compared with Experiment 1 in which there were 2 blocks of 6 lists. Second, immediately after the 12 study/interpolated task lists, participants were provided with a link that would be used to access a final recognition test that was to be completed 24–48 hr later.[2] Participants were not provided with a specific description of the final test at this time and the test could only be completed within the specified time frame to ensure a proper study-test delay. After the delay, all participants were provided with a 360-item recognition test, which was a combination of the recognition tests from the two blocks in Experiment 1. The test was randomized anew for each participant. Participants were compensated $2.50 for completion.

Results

Table 1 reports correct recall rates and false recall rates on the interpolated recall test and correct guessing proportions for critical items on the guessing task. Mean recognition scores are reported in Table 3. The data were analyzed as in Experiment 1.

Interpolated recall performance. Correct recall of list items varied across lists types, $F(2, 34) = 15.05$, $MSE = .004$, $\eta_p^2 = .47$. Correct recall was greater in categorized lists than both ad hoc (.58 vs. .54), $t(17) = 2.99$, $SEM = .01$, $d = 0.38$, and unrelated lists (.58 vs. .47), $t(17) = 4.13$, $SEM = .03$, $d = 0.95$. Recall for ad hoc lists was also greater than unrelated lists (.54 vs. .47), $t(17) = 3.77$, $SEM = .02$, $d = 0.58$.

Table 3
Mean (SD) Recognition Proportions for List Items and Critical Items of Categorized, Ad Hoc, and Unrelated Lists as a Function of Initial Testing Conditions for Experiment 2

Item type	Interpolated task group			
	Filler	Restudy	Recall	Guess
N	20	19	18	20
List items				
Categorized	.68 (.19)	.69 (.24)	.83 (.16)	.82 (.12)
Controls	.37 (.16)	.38 (.23)	.24 (.20)	.26 (.21)
Ad hoc	.56 (.18)	.63 (.23)	.66 (.15)	.65 (.19)
Controls	.35 (.17)	.33 (.22)	.24 (.16)	.23 (.15)
Unrelated	.45 (.19)	.51 (.25)	.54 (.18)	.49 (.19)
Controls	.31 (.18)	.29 (.19)	.24 (.16)	.24 (.19)
Corrected recognition				
Categorized	**.31 (.20)**	**.31 (.20)**	**.58 (.19)**	**.56 (.19)**
Ad hoc	**.20 (.18)**	**.30 (.22)**	**.42 (.18)**	**.43 (.18)**
Unrelated	**.14 (.15)**	**.22 (.13)**	**.30 (.14)**	**.25 (.15)**
Task average	**.22 (.15)**	**.27 (.16)**	**.43 (.14)**	**.41 (.15)**
Critical iems				
Categorized	.55 (.20)	.57 (.21)	.52 (.26)	.59 (.19)
Controls	.43 (.19)	.35 (.24)	.29 (.21)	.27 (.16)
Ad hoc	.49 (.16)	.44 (.22)	.38 (.23)	.44 (.15)
Controls	.37 (.19)	.37 (.23)	.32 (.23)	.24 (.19)
Unrelated	.31 (.19)	.32 (.22)	.22 (.19)	.19 (.20)
Controls	.28 (.21)	.27 (.23)	.20 (.18)	.22 (.19)
Corrected recognition				
Categorized	**.12 (.13)**	**.22 (.26)**	**.23 (.24)**	**.32 (.18)**
Ad hoc	**.13 (.15)**	**.07 (.20)**	**.06 (.22)**	**.20 (.18)**
Unrelated	**.03 (.16)**	**.05 (.12)**	**.02 (.13)**	**−.03 (.12)**
Task average	**.09 (.09)**	**.11 (.15)**	**.10 (.12)**	**.16 (.12)**

Note. Boldface indicates data used in the statistical analyses.

False recall, however, only differed marginally across list types, $F(2, 34) = 2.71$, $MSE = .001$, $p = .08$, $\eta_p^2 = .14$, $p_{BIC} = .82$. False recall was at floor following categorized and ad hoc lists and did not differ (.02 vs. .02), $t < 1$, $p_{BIC} = 1.00$, though false recall from these lists was greater than unrelated lists (.02 vs. .00), $t(17) = 2.92$, $SEM = .01$, $d = 1.02$, and (.02 vs. .00), $t(17) = 2.20$, $SEM = .01$, $d = 0.73$, respectively. Thus, as in Experiment 1, false recall of critical items was quite low across all list types.

Interpolated guessing performance. The proportion of correctly guessed critical items were analyzed as in Experiment 1 and were found to differ across list types, $F(2, 38) = 24.39$, $MSE = .01$, $\eta_p^2 = .56$. Successful guessing was greater on categorized lists than both ad hoc (.21 vs. .13), $t(19) = 3.04$, $SEM = .03$, $d = 0.62$, and unrelated lists (.21 vs. .00), $t(19) = 5.89$, $SEM = .04$, $d = 1.86$. Correct guessing was also greater on ad hoc than unrelated lists (.13 vs. .00), $t(19) = 4.80$, $SEM = .03$, $d = 1.52$. The amount of time spent guessing the critical items was 33.33 s per list, still less than the 60 s available in the retrieval-practice and the restudy groups.

Recognition. Recognition proportions were analyzed as in Experiment 1. Correct recognition differed across list types, $F(2, 146) = 60.60$, $MSE = .02$, $\eta_p^2 = .45$. Recognition was greater on categorized lists than both ad hoc lists (.44 vs. .33), $t(76) = 4.82$, $SEM = .02$, $d = 0.48$, and unrelated lists (.44 vs. .22), $t(76) = 10.04$, $SEM = .02$, $d = 1.10$. Recognition was also greater on ad hoc than unrelated lists (.33 vs. .22), $t(76) = 5.98$, $SEM = .02$, $d = 0.61$.

More important, correct recognition again differed across interpolated task groups, $F(3, 73) = 9.73$, $MSE = .07$, $\eta_p^2 = .29$. Recognition was greater in the retrieval-practice group than both the restudy group (.43 vs. .27), $t(35) = 3.31$, $SEM = .03$, $d = 1.12$, and the filler group (.43 vs. .22), $t(36) = 4.65$, $SEM = .03$, $d = 1.55$. More important, these retention benefits were once again found in the guess group which, like the retrieval-practice group, showed greater correct recognition relative to the restudy (.41 vs. .27), $t(37) = 2.82$, $SEM = .03$, $d = 0.93$, and filler groups (.41 vs. .22), $t(38) = 4.12$, $SEM = .03$, $d = 1.34$. As in Experiment 1, attempting to guess the nonpresented critical items was not more beneficial than retrieval-practice (43 vs. .41), $t < 1$, $p_{BIC} = .84$. Recognition was also equivalent between restudy and filler groups (.27 vs. .22), $t(37) = 1.14$, $SEM = .03$, $p = .27$, $p_{BIC} = .76$.

A significant List Type × Interpolated Task interaction was also found, $F(6, 146) = 3.76$, $MSE = .02$, $\eta_p^2 = .13$. Follow-up tests revealed that this interaction was related to greater increases in correct recognition after interpolated retrieval-practice and guessing tasks on categorized lists and smaller increases on ad hoc and unrelated lists. On categorized lists, both retrieval-practice and guessing groups reliably increased correct recognition relative to

[2] Because participants were provided with a window with which to complete the final recognition test, it is possible that interpolated task groups may have differed in the amount of time taken between study and recognition phases. Time differences would have influenced the retention interval which, in turn, may have contributed to recognition performance across groups. Mean retention intervals (in hours) were, therefore, compared across task groups. No differences in retention intervals were found between the filler ($M = 30.00$, $SD = 11.28$), restudy ($M = 31.92$, $SD = 6.72$), retrieval-practice ($M = 29.28$, $SD = 10.32$), or guess groups ($M = 32.40$, $SD = 9.12$), $F < 1$, suggesting that group differences are likely not attributable to retention interval differences.

filler and restudy groups, all ts > 4.04, ds > 1.31; however, on ad hoc and unrelated lists, retrieval-practice and guessing were only greater than the filler group, ts > 2.21, ds > 0.72, but not more so than the restudy group, ts < 1.98, ps $> .06$, ds < 0.64. Therefore, compared with the restudy group, the benefits of retrieval-practice and guessing tasks appear to diminish when the semantic relations of the list items are reduced.

False recognition was similarly analyzed as in correct recognition and was found to differ across list types, $F(2, 146) = 30.49$, $MSE = .03$, $\eta_p^2 = .30$. False recognition was greater on categorized lists than both ad hoc (.22 vs. .12), $t(76) = 3.88$, SEM .03, $d = 0.51$, and unrelated lists (.22 vs. .02), $t(76) = 7.40$, $SEM = .03$, $d = 1.13$. False recognition was also greater on ad hoc than unrelated lists (.12 vs. .02), $t(76) = 3.71$, $SEM = .03$, $d = 0.61$. False recognition did not differ across interpolated task groups, $F(3, 73) = 1.34$, $MSE = .04$, $p = .27$, $p_{BIC} = .99$; however, a List Type × Interpolated Task interaction was found, $F(6, 146) = 3.06$, $MSE = .03$, $\eta_p^2 = .11$. Follow-up t tests revealed that on categorized lists, guessing increased false recognition relative to the filler group (.32 vs. .12), $t(38) = 3.91$, $SEM = .04$, $d = 1.27$, but not more so than the restudy group (.32 vs. .22), $t(37) = 1.32$, $SEM = .05$, $p = .20$, $p_{BIC} = .72$, nor the retrieval-practice group (.32 vs. .23), $t(36) = 1.21$, $SEM = .05$, $p = .24$, $p_{BIC} = .74$. For ad hoc lists, false recognition in the guess group was greater than the restudy (.20 vs. .07), $t(37) = 2.10$, $SEM = .04$, $d = 0.69$, and retrieval-practice groups (.20 vs. .06), $t(36) = 2.16$, $SEM = .05$, $d = 0.72$, but only numerically greater than the filler group (.20 vs. .13), $t(38) = 1.39$, $SEM = .04$, $p = .17$, $p_{BIC} = .70$. For unrelated lists, however, all interpolated tasks were equivalent, ts < 1.97, ps $> .07$, p_{BIC}s $> .48$. Therefore, consistent with the ironic effect of guessing and Experiment 1, attempting to guess nonpresented critical items generally increased false recognition relative to other interpolated tasks, though this increase tended to be greater on ad hoc than categorized lists.

Discussion

Retrieval-practice and guess interpolated task performance was again sensitive to the semantic structure of the study list. Correct recall of list items and correct guessing of critical items was greatest for categorized lists and least for unrelated lists. Similar to Experiment 1, correct recognition on categorized lists was greater after retrieval-practice and guessing interpolated tasks than the restudy and filler tasks; however, on ad hoc and unrelated lists, retrieval-practice and guessing tasks were only greater than the filler group. Across all list types, retrieval-practice and guess groups again showed similar correct recognition rates, suggesting that the guess task is not more susceptible to forgetting than interpolated recall. The guessing benefit was similar to recall despite the guess group using about half the amount of time attempting to guess items that were not presented than the recall group who recalled items that were presented.

For critical items, interpolated false recall was again at floor, but greater on categorized and ad hoc lists than unrelated lists. On the final recognition test, critical item false recognition was greatest on categorized lists followed by ad hoc and unrelated lists. False recognition was also influenced by the type of interpolated task. Consistent with the ironic effect of guessing, false recognition of categorized and ad hoc critical items was greatest in the guess group. Therefore, the correct recognition benefit in the interpolated guessing group was again accompanied by a concomitant increase in critical item false recognition for categorized and ad hoc critical items on a delayed test. As in Experiment 1, although the guess group showed some success at identifying critical items during the guessing phase on categorized lists, this came at the cost of greater false recognition. Successful guessing of categorized critical items was again fairly low (21%), which may not have been sufficient to reduce later false recognition. When both correct and false recognition are considered, interpolated guessing again appears to produce both benefits and costs when compared with a restudy control group.

Experiment 3: Within-Subjects Interpolated Tasks

The previous two experiments have shown that attempting to guess a set of nonpresented critical items after study of a word list enhances subsequent recognition of those list items akin to retrieval practice. When considered in the context of cue-target guessing experiments (e.g., Grimaldi & Karpicke, 2012; Hays et al., 2013; Huelser & Metcalfe, 2012; Knight et al., 2012; Kornell et al., 2009; Potts & Shanks, 2014; Yan et al., 2014), information used to generate a guess also appears to show a memory benefit over interpolated restudy and filler tasks. Although our experiments have shown that guessing benefits occur both when final recognition testing is completed immediately and after a delay, the mechanism(s) underlying the guessing benefit is less well understood. One possibility, as mentioned above, is that guessing enhances retention by operating as an implicit retrieval process where participants mentally retrieve study items to generate critical items. Such retrieval processes have also been proposed as a major mechanism underlying testing effect benefits (see Kornell, Bjork, & Garcia, 2011, for a bifurcation account of retrieval practice). The only difference is that in the guessing condition, individuals are implicitly retrieving items after list presentation in search of how they may be related to a critical nonpresented item. Another possibility, however, is that guessing operates during the list presentation to shape how participants process the word list as the items are presented. In Experiments 1 and 2 (and Huff et al., 2012), the guess group is presented with guessing instructions *before* the presentation of the first list (indeed, the filler, restudy, and retrieval-practice groups all knew what test to expect before the list presentations). Because participants were informed of the upcoming task requirements, they may relate items during study to generate critical items, rather than by implicitly retrieving the items after the list is presented. In other words, guessing benefits may not be driven by the interpolated guess task per se, but instead by preparatory encoding processes used to generate a set of plausible critical items.

Previous research has shown that the expectation of an upcoming memory test can affect how a memory representation is formed at study. For example, Neely and Balota (1981; see too Balota & Neely, 1980) have shown that expectations of an upcoming free recall test lead to an increase in processing of study items (i.e., node tagging; Anderson & Bower, 1972) relative to expectations of an upcoming recognition test. This pattern is noteworthy because the expectation of a recall test does not provide an explicit encoding strategy for participants to utilize, though participants still modify their encoding in preparation for the upcoming test.

Given these test-expectancy processes, it is not unreasonable to suspect that participants in the interpolated-guess group are also tuning their encoding strategy to identify the critical items. This is especially likely because the guess group is explicitly instructed that the list items are related to each other along some dimension, and they were to guess five missing words.

Additional evidence for relational encoding in the guessing group can be found in critical item false recognition. In both previous experiments, false recognition was generally greatest in the guess group relative to the other interpolated tasks—a pattern that parallels relational encoding tasks in the DRM paradigm. Specifically, tasks that direct encoding toward relations shared among list items typically increase the false memory illusion relative to tasks that direct encoding to the unique features of each list item (item-specific encoding; Huff & Bodner, 2013; McCabe, Presmanes, Robertson, & Smith, 2004). Thus, as participants relate the list items to identify critical items at study, this processing may enhance correct recognition, but also produce a cost by increasing false recognition.

The purpose of Experiment 3 is therefore to evaluate whether the guessing benefit shown in the previous two experiments is because of processes in operation during encoding or during the interpolated task. A simple method to reduce preparatory encoding processes, or at the very least make them more similar across lists, is to manipulate encoding instructions randomly within subjects such that participants are unaware of the exact task that will be completed during study. If participants are unsure of the type of interpolated task they will complete after study, they will be unable to recruit specific encoding processes in preparation for the upcoming task. Therefore, if guess participants are using more relational processing during encoding to generate critical items, they will be unlikely to do so under random instruction conditions that should result in similar processing across lists. Further, by presenting tasks randomly after study, we can also evaluate how preparatory encoding may be contributing to retrieval-practice benefits in the recall group.

In Experiment 3, all participants were presented with restudy, recall, and guess instructions before presentation of the study lists, so they understood the demands of each task before study. Participants were further instructed that, after each list, one of the three tasks will be completed randomly, but the specific task would not be revealed until after each list was studied. Interpolated tasks were therefore completed randomly after each list. A final recognition test was completed after all lists were studied. If the guessing benefit in the previous experiments is because of preparatory encoding processes, then final recognition on guess lists would be similar to that on the restudy lists. In contrast, if guessing still shows a benefit over restudy, then this benefit is likely because of covert retrieval processes during the guessing task. Similar predictions were made for the retrieval-practice group in which encoding based processes may contribute to the benefits of retrieval-practice. Because retrieval practice still occurs after list presentation in the recall test group, one should still find benefits of retrieval practice on recognition relative to the restudy group.

Method

Participants. Thirty-three individuals were recruited using Amazon's Mechanical Turk in this within-subject's design. All were proficient in English and resided in the United States with normal or corrected-to-normal vision. Mean reported age was 39.30 years (SD = 13.67, range = 20–68) and mean reported formal education was 15.63 years (SD = 1.67, range = 12–19). One participant reported cheating on the final questionnaire and two participants recalled items during lists with guessing instructions and were not included leaving 30 participants for analysis.

Materials and procedure. All procedures and stimulus lists used in Experiment 3 were identical to those of Experiment 2 with the following exceptions. First, the filler task condition was eliminated given the similarities to the restudy condition in correct and false recognition in the previous experiments. Second, interpolated task instructions were manipulated within-subjects. Before studying the first list, all participants were provided with an instruction screen that displayed instructions for each of the restudy, recall, and guessing tasks. Specifically, participants were informed that after the presentation of each list, they would complete one of three tasks. For the *restudy task*, participants were told that they would restudy the words they just saw but in a different order. For the *recall task*, participants were told that they would complete a memory task on the words that were just presented by typing them into a dialogue box for 1 min. In the *guess task*, participants were told that five words were related to the list they just saw in some way but were not actually presented in the list itself and would attempt to guess up to five of these nonpresented words (and at least one), by typing them into a dialogue box. Task instructions for all three tasks were presented to all participants on the same screen. Critically, participants were also told that the task they will complete will occur randomly and they will be informed of the specific task only after the list was presented. Therefore, they would not be aware of the specific interpolated task they would later complete at the time of study. Instructions for each interpolated task were again presented after study of each list to cue participants to complete the appropriate interpolated task.

Three separate versions were created that counterbalanced the order of the interpolated tasks as well as the list type across participants. The total number of lists studied was reduced to nine lists so that each list type (categorized, ad hoc, and unrelated) was studied before completing one of the three interpolated tasks. The recognition test was again completed after a 24–48 hr delay as in Experiment 2, and contained 270 total items that were randomized anew for each participant. Participants were compensated $2.50 for participation.

Results

Table 1 displays correct recall rates and false recall rates on the interpolated recall test and correct guessing proportions for critical items on the guessing task. Mean recognition scores are reported in Table 4. Data were analyzed as in Experiment 1.

Interpolated recall performance. Correct recall was again found to vary as a function of list type, $F(2, 58) = 12.22$, $MSE = .02$, $\eta_p^2 = .30$; however, recall was only numerically greater on categorized than ad hoc lists (.43 vs. .42), $t < 1$, $p_{BIC} = .82$, but significantly greater on categorized than unrelated lists (.43 vs. .29), $t(29) = 3.86$, $SEM = .04$, $d = 0.78$. Recall was also greater on ad hoc than unrelated lists (.41 vs. .29), $t(29) = 4.18$, $SEM = .03$, $d = 0.61$.

Table 4
Mean (SD) Recognition Proportions for List Items and Critical Items of Categorized, Ad Hoc, and Unrelated Lists as a Function of Interpolated Task Lists for Experiment 3

Item type	Interpolated task		
	Restudy	Recall	Guess
List items			
Categorized	.82 (.21)	.80 (.20)	.83 (.19)
Controls		.36 (.21)	
Ad hoc	.73 (.19)	.70 (.20)	.60 (.26)
Controls		.29 (.20)	
Unrelated	.59 (.22)	.56 (.22)	.58 (.19)
Controls		.27 (.18)	
Corrected recognition			
Categorized	**.46 (.28)**	**.45 (.25)**	**.47 (.25)**
Ad hoc	**.44 (.28)**	**.41 (.26)**	**.31 (.33)**
Unrelated	**.32 (.28)**	**.29 (.25)**	**.31 (.21)**
Task average	**.41 (.23)**	**.38 (.19)**	**.37 (.22)**
Critical items			
Categorized	.47 (.31)	.64 (.35)	.42 (.34)
Controls		.34 (.27)	
Ad hoc	.46 (.30)	.52 (.26)	.50 (.35)
Controls		.33 (.24)	
Unrelated	.40 (.27)	.29 (.27)	.37 (.31)
Controls		.30 (.20)	
Corrected recognition			
Categorized	**.13 (.32)**	**.30 (.36)**	**.08 (.32)**
Ad hoc	**.13 (.30)**	**.19 (.24)**	**.17 (.33)**
Unrelated	**.10 (.22)**	**−.01 (.19)**	**.07 (.22)**
Task average	**.11 (.16)**	**.16 (.17)**	**.11 (.16)**

Note. $N = 30$. Boldface indicates data used in the statistical analyses.

False recall was also found to differ across list types, $F(2, 58) = 5.42$, $MSE = .004$, $\eta_p^2 = .16$, with false recall greater on categorized lists than ad hoc (.05 vs. .01), $t(29) = 2.26$, $SEM = .02$, $d = 0.44$, and unrelated lists (.05 vs. .00), $t(29) = 2.50$, $SEM = .02$, $d = 0.65$. False recall did not differ on ad hoc and unrelated lists (.01 vs. .00), $t(29) = 1.44$, $SEM = .01$, $p = .16$, $p_{BIC} = .63$, but like Experiments 1 and 2, false recall was quite low across all list types.

Interpolated guessing performance. Correct guessing of nonpresented critical items was also found to differ across list types, $F(2, 58) = 15.38$, $MSE = .03$, $\eta_p^2 = .35$. Correct guessing was greater on categorized lists than both ad hoc (.23 vs. .05), $t(29) = 3.53$, $SEM = .05$, $d = 0.85$, and unrelated lists (.23 vs. .00), $t(29) = 4.75$, $SEM = .05$, $d = 1.23$. Correct guessing was only marginally greater on ad hoc than unrelated lists, $t(29) = 1.76$, $SEM = .03$, $p = .09$, $d = 0.45$, $p_{BIC} = .54$. Thus, successful guessing was once again moderated by the relatedness of the studied list. The average amount of time per list spent guessing the critical lures was 34.06 s.

Recognition. To determine whether randomly presented interpolated tasks affected final correct recognition, a 3 (List Type) × 3 (Interpolated Task) within-subjects ANOVA was conducted. Correct recognition was found to vary across list types, $F(2, 116) = 10.14$, $MSE = .03$, $\eta_p^2 = .26$. Correct recognition was only marginally greater on categorized than ad hoc lists (.46 vs. .39), $t(29) = 1.90$, $SEM = .04$, $p = .07$, $d = 0.30$, $p_{BIC} = .62$, but significantly greater on categorized than unrelated lists (.46 vs. .30), $t(29) = 4.60$, $SEM = .03$, $d = 0.72$. Correct recognition was also greater on ad hoc than unrelated lists (.39 vs. .30), $t(29) = 2.66$, $SEM = .03$, $d = 0.35$. Critically, both the main effect of Interpolated Task and the interaction failed to reach significance, Fs < 1.55, ps $> .19$, p_{BIC}s $> .72$, demonstrating that varying interpolated tasks randomly after the presentation of each list, thereby eliminating task-specific encoding operations including the guessing benefit and the retrieval-practice effect in correct recognition.

Turning to false recognition of critical items, an effect of List Type was once again found, $F(2, 116) = 3.38$, $MSE = .06$, $\eta_p^2 = .10$. False recognition was equivalent between categorized and ad hoc lists (.17 vs. .16), $t < 1$, $p_{BIC} = .85$, but greater on categorized than unrelated lists (.17 vs. .05), $t(29) = 2.07$, $SEM = .06$, $d = 0.56$, and greater on ad hoc than unrelated lists (.16 vs. .05), $t(29) = 3.15$, $SEM = .04$, $d = 0.67$. The effect of Interpolated Task was not significant, $F(2, 116) = 1.26$, $MSE = .06$, $p = .29$, $p_{BIC} = .75$; however, a List Type × Interpolated Task interaction was found, $F(4, 116) = 3.82$, $MSE = .06$, $\eta_p^2 = .12$. Follow-up tests revealed that this interaction was related to elevated false recognition after recall relative to other interpolated tasks but selectively on categorized lists. Specifically, categorized false recognition was greater after recall than restudy (.30 vs. .13), $t(29) = 2.49$, $SEM = .07$, $d = 0.50$, and after recall than guessing (.30 vs. .08), $t(29) = 3.60$, $SEM = .06$, $d = 0.64$. False recognition was equivalent after restudy and guessing (.13 vs. .08), $t < 1$, $p_{BIC} = .80$, showing an elimination of the ironic effect of guessing when the interpolated task occurred randomly after study. False recognition was marginally greater after restudy than recall on unrelated lists (.10 vs. −.01), $t(29) = 2.04$, $SEM = .05$, $p = .05$, $d = 0.53$, $p_{BIC} = .42$, and all other comparisons were not significant, ts < 1.63, ps $> .12$, p_{BIC}s $= .89$.

Discussion

The primary finding in Experiment 3 was the elimination of the guessing interpolated task benefit on correct recognition across all list types when tasks were completed randomly after study. We suggest that random task presentation eliminates task specific preparatory encoding processes, and hence the guessing benefits found in Experiments 1 and 2 (and Huff et al., 2012) were at least in part because of encoding processes.

Also noteworthy, retrieval-practice benefits over restudy were also eliminated under random instructions. Like guessing, this pattern suggests that retrieval-practice effects, at least those that occur when feedback is not provided, may in part be because of encoding processes rather than retrieval-based processes at test. This pattern is surprising, given the robust benefits shown by retrieval practice paradigms (cf. Rawson & Dunlosky, 2011); however, previous work has indeed shown that the magnitude of retrieval-practice benefits can be moderated based on whether a final memory test is expected or not. It is possible that the expectation of a retrieval task influences the processes engaged in during encoding, and so contributes to the benefits of testing. For instance, Szpunar, McDermott, and Roediger (2007) showed that participants who expected a final cumulative test on a series of five study lists showed a larger retrieval-practice effect on final recall than participants who did not expect a final cumulative test. In addition, participants who are given intentional retrieval instructions to retrieve specific information are more likely to provide

recollection judgments on a final test than participants who are given incidental instructions (Pu & Tse, 2014). Thus, the expectation of recall may lead participants to process information more deeply at study, which may contribute to the retrieval-practice benefits observed in Experiments 1 and 2.

For critical item false recognition, the ironic effect of guessing on categorized and ad hoc lists was similarly eliminated under random task conditions, providing additional support for the contribution of preparatory processes in the observed guessing effects observed in Experiments 1 and 2. In contrast, interpolated recall tasks did show increased false recognition for categorized lists, demonstrating that, even when test expectancy processes are minimized at study, retrieval practice via free recall can inflate false recognition.

General Discussion

The purpose of the present experiments was to evaluate the effects of interpolated guessing, retrieval practice, restudy, and filler tasks on final correct recognition of list items and false recognition of nonpresented critical items on categorized, ad hoc, and unrelated list types. We replicated Huff et al.'s (2012) findings that attempting to guess nonstudied items improves final correct recognition relative to a filler task and importantly, showed this benefit persists relative to a restudy control group—a critical comparison not included in this earlier work. This pattern of results shows a memory benefit for information that is used to generate a guess (cf. Kornell et al., 2009). Further, we showed that interpolated guessing can also result in a cost of greater critical item false recognition, consistent with the ironic effect of guessing (Huff et al., 2012). The benefits and costs of interpolated guessing were extended to when final recognition was completed both within the same experimental session as list study (Experiment 1) and when final recognition was delayed at least 24 hr (Experiment 2).

In Experiment 3, we sought to determine if interpolated guessing benefits were because of a task-expectancy process, given that participants in prior experiments were presented with instructions regarding the type of task they would receive before study and that participants completed the same interpolated task after each study list. Task type was manipulated within subjects in Experiment 3, with task instructions presented randomly after study to minimize expectancy effects. Unlike Experiments 1 and 2, both correct and false recognition in Experiment 3 for guess lists was similar to that of restudy lists, suggesting that previous guessing patterns were likely because of task-expectancy-based encoding processes.

An additional goal of our experiments was to evaluate whether interpolated guessing produced similar memory benefits to those of retrieval practice given previous work showing that guessing may produce even stronger memory benefits than recall. To this end, we compared a retrieval-practice group to the restudy and guess groups to gauge guessing benefits within the context of a retrieval-practice paradigm (cf. Roediger & Karpicke, 2006). If guessing operates to improve retention, it is important to determine whether correct recognition benefits are at least as potent as those after retrieval practice. Indeed, correct recognition after interpolated guessing was greater than restudy in Experiments 1 and 2 and equivalent to the retrieval-practice group. Further, to provide an additional control comparison group, we included a filler task group that was not re-exposed to the study items. Guessing and retrieval practice were also greater than the filler group, thus with both types of controls, guessing produced a benefit on correct recognition.

Although the guessing task in Experiments 1 and 2 produced greater correct recognition over controls, it never exceeded the level of the retrieval-practice group. This outcome is curious given that Huff et al.'s (2012) interpolated guess group outperformed interpolated retrieval practice in correct recognition. One possibility for this difference could be related to differences in the relations between items in the study lists. For instance, in Huff et al., study items shared either direct or indirect semantic associations to a nonpresented critical item. In contrast, list items used in the present experiments were either unrelated to critical items or were related to critical items though a shared superordinate category rather than a direct or indirect associative relationship. For example, the critical item "Blue Jay" does not share an associative relationship to "Oriole" according to free association norms (Nelson, McEvoy, & Schreiber, 1999), but instead the two items share a categorical relationship. It is possible that guessing produces a greater memory improvement when study information shares semantic associations rather than categorical relations. This possibility is particularly likely if participants are engaging in task-expectancy processes such as relational processing, which is likely the default processing used when list items are related (Huff & Bodner, 2014; Hunt & Seta, 1984).

Given a similar correct recognition benefit between the guessing and retrieval-practice groups in Experiments 1 and 2, we note that guessing of information that was not studied could be an effective and efficient method of improving subsequent recognition. Participants in the guess groups spent considerably less time guessing items that were not studied than participants who retrieved items that were studied. When considered with previous guessing experiments that provide corrective feedback (e.g., Kornell et al., 2009; Potts & Shanks, 2014), guessing may be a powerful study method that can be applied in educational settings given that it improves memory for the information used to generate a guess, and improves memory for corrective feedback as in the previous literature. However, guessing can also produce a cost by increasing false alarms to related lures that may limit the utility of guessing instructions if the final test contains related lures.

Our study also provides evidence that the expectation of an upcoming task can shape how individuals process information in preparation for the upcoming task. This is particularly likely under guessing instructions, as participants are instructed to generate a set of critical items that are related to the study list. A marker of relational encoding is an increase in false recognition for related lures (Huff & Bodner, 2013) and consistent with this pattern, guessing produced the highest false recognition rate in Experiments 1 and 2. In Experiment 3, task expectancy was minimized because of the random task presentation after study. As a result, guessing effects on both correct and false recognition were eliminated relative to restudy, providing additional evidence that guessing instructions operate to increase preparatory relational encoding.

Interestingly, the random presentation of task instructions after study similarly eliminated the retrieval-practice benefit over restudy on the final recognition test. This pattern suggests that retrieval-practice benefits on recognition performance, at least without feedback, may in part be because of the expectancy of an

upcoming recall test. As mentioned above, other studies have shown that retrieval-practice effects are either larger or accompanied by more detailed retrievals when participants have retrieval expectations at study (e.g., Pu & Tse, 2014; Szpunar et al., 2007). Given the educational applications of retrieval practice, an important area for future research is to determine the extent to which retrieval-practice effects are driven by expectancy-based processes that may affect how materials are studied or if benefits are due purely to retrieval-based processes.

Accounts of retrieval practice postulate that testing operates to increase the number of available retrieval routes that can be used on subsequent tests (Bjork, 1975; Roediger & Karpicke, 2006) or through the generation of semantic mediators that act as retrieval cues (Carpenter, 2011; Pyc & Rawson, 2010), both of which do not address how expectancy effects may shape processing of the list during study. Our experiment suggests that the production of retrieval routes/cues may occur at least in part during study in anticipation of an upcoming test rather than during retrieval, provided a recall test is expected. That is, the retrieval practice group may anticipate an upcoming recall test that may facilitate encoding processes (see Balota & Neely, 1980) relative to the restudy group in which participants know they will be represented with the study information. In the case of restudy, it is possible that participants may "loaf" during the first study phase knowing that they will be presented with this information again, whereas participants who are aware of an upcoming test may process study information more deeply to maximize test performance. Given the utility of retrieval practice both in basic and applied research, it is, therefore, critical to understand the role of test expectancy in contributing to testing effects. For instance, it is important to determine whether it is necessary for students to encode educational materials in preparation for an upcoming test, only complete a memory test, or expect and complete a memory test to maximize retention. Of course, the primary focus of our experiments was on the effects of guessing on subsequent recognition. However, we note the possibility that contributions of test expectancy in retrieval-practice paradigms may be underestimated as participants are often aware that they will either restudy or be tested on a study set.

An additional possibility for the lack of a retrieval-practice benefit in Experiment 3 may be because of how well participants are able to successfully retrieve items on an initial test. As can be seen in Table 1, correct recall in Experiment 3 was between .12 and .18 lower than correct recall in Experiment 2 where a robust retrieval-practice benefit was found. It is possible that correct recall in Experiment 3 was not sufficiently high to produce a retrieval-practice advantage over restudy. To address this possibility, we conducted a median-split analysis that divided participants into low and high recall groups based on correct recall during the interpolated task. Correct recall was .26 in the low recall group, whereas correct recall was .53 in the high recall group—a rate similar to Experiments 1 and 2. A reanalysis of the corrected recognition data with low and high recall groups as a between-subjects factor showed no interaction with either interpolated task type or list type, Fs < 1, demonstrating no task differences regardless of whether initial correct recall performance was relatively low or relatively high. Thus, the lack of a retrieval-practice effect in Experiment 3 does not appear to be related to a relatively low correct recall rate on the interpolated task.

It is also important to emphasize that, although we argue that differential task-expectancy processes are minimized because of random instructions after study in Experiment 3, this pattern covaries with a change to a within-subjects design that precludes our study from separating expectancy versus design contributions. Thus, the equivalent task patterns reported could be because of similar task expectancies, the use of a within design, or some combination of the two. While determining the contributions of experimental design and expectancy because of the timing of instructions is important, it is worth noting that establishing sufficient task expectancies to alter encoding may be more complex than simply exposing participants to task instructions before study. For example, Neely and Balota (1981) had participants complete six practice study-test cycles to show test expectancy processing effects on a critical list. Similarly, in Experiments 1 and 2, participants completed 6 and 12 consecutive study-task cycles, respectively, which may have been necessary to establish expectancy processing. Therefore, the amount of task practice may also be an important determinant of task-expectancy effects.

One method for examining the development of expectancy effects over study-task cycles—a suggestion provided by an anonymous reviewer—is to compare task effects in Experiment 1 where participants completed a recognition test separately on two blocks of six lists. If expectancy effects require exposure to several study-task cycles to develop, one would expect larger task-expectancy effects for the second block of lists than the first. Consistent with this rationale, a reanalysis of corrected recognition of list items in Experiment 1 with recognition block entered as a factor revealed a significant Block by Interpolated Task interaction, $F(3, 304) = 3.13$, $MSE = .02$, $p = .03$, in which the size of the retrieval-practice benefit over restudy was indeed larger on Block 2 than Block 1 (.18 vs. .08), as was the size of the guessing benefit (.12 vs. .07). An analysis of corrected false recognition yielded a marginal Block by Interpolated Task interaction, $F(3, 304) = 2.22$, $MSE = .03$, $p = .09$, which similarly showed larger recall and guessing effects on Block 2 than Block 1. The three-way interaction with List Type for both correct and false recognition was not significant, $F < 1$, showing similar patterns across list types. Therefore, interpolated task benefits and costs are found for both blocks when participants have task expectancies, but their effects do appear to increase in magnitude as study-task cycles repeat, which is consistent with our contention that expectancy effects do increase over repeated task exposure. It is still unclear, however, whether one or many task repetitions are necessary to produce task effects, but this analysis does suggest that greater repetitions may facilitate anticipatory task-based processing.

Conclusions

The present study was designed to evaluate the extent to which attempting to guess a set of nonpresented critical items from a word list influences subsequent memory for the words presented on that list. The results of this study were noteworthy in that attempting to guess critical items did produce a correct recognition benefit relative to restudy of that list. However, when the interpolated task was not known until after the study list was presented—a procedure designed to reduce expectancy-based processes—the recognition benefit found after guessing was eliminated. Guessing benefits appear to operate by enhancing expectancy processes that

shape how participants study the list. In addition, this study demonstrated that task expectancy may also contribute to retrieval-practice benefits as the recognition benefit found after the completion of a recall test was similarly eliminated when expectancy for recall were diminished. Therefore, the present experiments add to previous work showing that guessing and retrieval-practice can be powerful methods for improving memory performance, but that these benefits may in part be because of expectancy-based encoding processes, as opposed to retrieval-based processes after the list is presented. Future research should be mindful of whether task-induced effects are attributable to task-based processes or expectancy driving participant's preparation for that task.

References

Anderson, J. R., & Bower, G. H. (1972). Recognition and retrieval processes in free recall. *Psychological Review, 79,* 97–123. http://dx.doi.org/10.1037/h0033773

Baddeley, A., & Wilson, B. A. (1994). When implicit learning fails: Amnesia and the problem of error elimination. *Neuropsychologia, 32,* 53–68. http://dx.doi.org/10.1016/0028-3932(94)90068-X

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. http://dx.doi.org/10.1037/0278-7393.6.5.576

Balota, D. A., Yap, M. J., Cortese, M. J., Hutchison, K. A., Kessler, B., Loftis, B., . . . Treiman, R. (2007). The English lexicon project. *Behavior Research Methods, 39,* 445–459. http://dx.doi.org/10.3758/BF03193014

Bartlett, S. F. C. (1932/1967). *Remembering: A study in experimental and social psychology*. Cambridge, United Kingdom: 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. http://dx.doi.org/10.1037/h0027577

Bengson, J. J., & Hutchison, K. A. (2007). Variability in response criteria affects estimates of conscious identification and unconscious semantic priming. *Consciousness and Cognition, 16,* 785–796. http://dx.doi.org/10.1016/j.concog.2006.12.002

Benjamin, A. S. (2001). On the dual effects of repetition on false recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27,* 941–947. http://dx.doi.org/10.1037/0278-7393.27.4.941

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.

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. http://dx.doi.org/10.1037/a0024140

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

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

Craik, F. I. M. (2002). Levels of processing: Past, present, and future? *Memory, 10,* 305–318. http://dx.doi.org/10.1080/09658210244000135

Craik, F. I. M., & Lockhart, R. S. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior, 11,* 671–684. http://dx.doi.org/10.1016/S0022-5371(72)80001-X

Deese, J. (1959). On the prediction of occurrence of particular verbal intrusions in immediate recall. *Journal of Experimental Psychology, 58,* 17–22. http://dx.doi.org/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. http://dx.doi.org/10.1037/0278-7393.30.1.120

Grimaldi, P. J., & Karpicke, J. D. (2012). When and why do retrieval attempts enhance subsequent encoding? *Memory & Cognition, 40,* 505–513. http://dx.doi.org/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. http://dx.doi.org/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. http://dx.doi.org/10.3758/s13421-011-0167-z

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. http://dx.doi.org/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. http://dx.doi.org/10.1016/j.jml.2014.02.004

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, 38,* 1720–1730. http://dx.doi.org/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. http://dx.doi.org/10.3758/s13421-013-0299-4

Huff, M. J., & Hutchison, K. A. (2011). The effects of mediated word lists on false recall and recognition. *Memory & Cognition, 39,* 941–953. http://dx.doi.org/10.3758/s13421-011-0077-0

Huff, M. J., Meade, M. L., & Hutchison, K. A. (2011). Age-related differences in guessing on free and forced recall tests. *Memory, 19,* 317–330. http://dx.doi.org/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. http://dx.doi.org/10.1016/S0022-5371(81)90138-9

Hunt, R. R., & Seta, C. E. (1984). Category size effects in recall: The roles of relational and individual item information. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 10,* 454–464. http://dx.doi.org/10.1037/0278-7393.10.3.454

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. http://dx.doi.org/10.1037/a0021977

Kantner, J., & Lindsay, D. S. (2012). Response bias in recognition memory as a cognitive trait. *Memory & Cognition, 40,* 1163–1177. http://dx.doi.org/10.3758/s13421-012-0226-0

Kay, H. (1955). Learning and retaining verbal material. *British Journal of Psychology, 46,* 81–100. http://dx.doi.org/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. http://dx.doi.org/10.1016/S0749-596X(02)00504-1

Knight, J. B., Ball, B. H., Brewer, G. A., DeWitt, M. R., & Marsh, R. L. (2012). Testing unsuccessfully: A specification of the underlying mech-

anisms supporting its influence on retention. *Journal of Memory and Language, 66,* 731–746. http://dx.doi.org/10.1016/j.jml.2011.12.008

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

Kornell, N., Bjork, R. A., & Garcia, M. A. (2011). Why tests appear to prevent forgetting: A distribution-based bifurcation model. *Journal of Memory and Language, 65,* 85–97. http://dx.doi.org/10.1016/j.jml.2011.04.002

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. http://dx.doi.org/10.1037/a0015729

Lane, S. M., Mather, M., Villa, D., & Morita, S. K. (2001). How events are reviewed matters: Effects of varied focus on eyewitness suggestibility. *Memory & Cognition, 29,* 940–947. http://dx.doi.org/10.3758/BF03195756

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

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

Masson, M. E. J. (2011). A tutorial on a practical Bayesian alternative to null-hypothesis significance testing. *Behavior Research Methods, 43,* 679–690. http://dx.doi.org/10.3758/s13428-010-0049-5

McCabe, D. P., Presmanes, A. G., Robertson, C. L., & Smith, A. D. (2004). Item-specific processing reduces false memories. *Psychonomic Bulletin & Review, 11,* 1074–1079. http://dx.doi.org/10.3758/BF03196739

McDermott, K. B. (1996). The persistence of false memories in list recall. *Journal of Memory and Language, 35,* 212–230. http://dx.doi.org/10.1006/jmla.1996.0012

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

Meade, M. L., & Roediger, H. L., III. (2009). Age differences in collaborative memory: The role of retrieval manipulations. *Memory & Cognition, 37,* 962–975. http://dx.doi.org/10.3758/MC.37.7.962

Nairne, J. S., Thompson, S. R., & Pandeirada, J. N. S. (2007). Adaptive memory: Survival processing enhances retention. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33,* 263–273. http://dx.doi.org/10.1037/0278-7393.33.2.263

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

Nelson, D. L., McEvoy, C. L., & Schreiber, T. (1999). *The University of South Florida Free Association Norms*. Retrieved from http://w3.usf.edu/FreeAssociation/

Potts, R., & Shanks, D. R. (2014). The benefit of generating errors during learning. *Journal of Experimental Psychology: General, 143,* 644–667. http://dx.doi.org/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. http://dx.doi.org/10.1007/s10339-013-0580-2

Pyc, M. A., & Rawson, K. A. (2010). Why testing improves memory: Mediator effectiveness hypothesis. *Science, 330,* 335. http://dx.doi.org/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. http://dx.doi.org/10.1037/a0023956

Roediger, H. L., III, Jacoby, J. D., & McDermott, K. B. (1996). Misinformation effects in recall: Creating false memories through repeated retrieval. *Journal of Memory and Language, 35,* 300–318. http://dx.doi.org/10.1006/jmla.1996.0017

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

Roediger, H. L., III, & McDermott, K. B. (1995). Creating false memories: Remembering words not presented in lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21,* 803–814. http://dx.doi.org/10.1037/0278-7393.21.4.803

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

Roediger, H. L., III, 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.

Seamon, J. G., Luo, C. R., Schwartz, M. A., Jones, K. J., Lee, D. M., & Jones, S. J. (2002). Repetition can have similar or different effects on accurate and false recognition. *Journal of Memory and Language, 46,* 323–340. http://dx.doi.org/10.1006/jmla.2001.2811

Skinner, B. F. (1953). *Science and human behavior*. Oxford, England: Macmillan.

Skinner, B. F. (1958). Teaching machines; from the experimental study of learning come devices which arrange optimal conditions for self instruction. *Science, 128,* 969–977. http://dx.doi.org/10.1126/science.128.3330.969

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

Snodgrass, J. G., & Corwin, J. (1988). Pragmatics of measuring recognition memory: Applications to dementia and amnesia. *Journal of Experimental Psychology: General, 117,* 34–50. http://dx.doi.org/10.1037/0096-3445.117.1.34

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

Tse, C.-S., Balota, D. A., & Roediger, H. L., III. (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. http://dx.doi.org/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. http://dx.doi.org/10.1016/j.jml.2003.10.003

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of *p* values. *Psychonomic Bulletin & Review, 14,* 779–804. http://dx.doi.org/10.3758/BF03194105

Wickens, T. D. (2002). *Elementary signal detection theory*. New York, NY: Oxford University Press.

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. http://dx.doi.org/10.3758/s13421-014-0454-6

Received September 30, 2015
Revision received January 29, 2016
Accepted January 31, 2016 ■