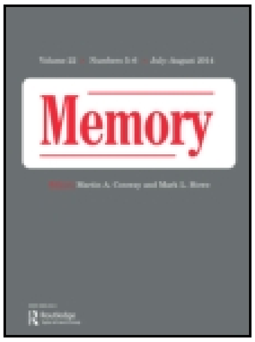

Memory

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

Adapting to test structure: Letting testing teach what to learn

Leonel Garcia-Marques, Ludmila D. Nunes, Pedro Marques, Paula Carneiro & Yana Weinstein

Memory, 2015
Vol. 23, No. 3, 365–380, http://dx.doi.org/10.1080/09658211.2014.889166

Routledge
Taylor & Francis Group

Adapting to test structure: Letting testing teach what to learn

Leonel Garcia-Marques[1], Ludmila D. Nunes[1,2], Pedro Marques[1], Paula Carneiro[1], and Yana Weinstein[3]

[1]Faculty of Psychology, University of Lisbon, Lisbon, Portugal
[2]Department of Psychological Sciences, Purdue University, West Lafayette, IN, USA
[3]Department of Psychology, University of Massachusetts – Lowell, Lowell, MA, USA

(*Received 23 October 2013; accepted 24 January 2014*)

We propose that we encode and store information as a function of the particular ways we have used similar information in the past. More specifically, we contend that the experience of retrieval can serve as a powerful cue to the most effective ways to encode similar information in comparable future learning episodes. To explore these ideas, we did two studies in which all participants went through study–test cycles of single category lists while we manipulated the nature of the recognition tests. The recognition tests either included only same-category lures or only different-category lures. The experience of repeated testing leads participants to avoid conceptual-based strategies but only when conceptual knowledge was poorly diagnostic for recognition (i.e., in the same-category lures condition). In a second study with a similar manipulation, we showed that repeated testing with lures from the same category as study items improved performance in a final recall surprise test compared to conditions in which different-category lures were used. Such a difference is akin to the one obtained when encoding instructions focus on distinctive item features compared to cases in which the focus is on relational processing. We suggest that testing requirements lead to adaptive changes at encoding.

Keywords: Retrieval practice; Implicit learning; Test structure; Adaptive memory.

> Human memory did not evolve to give optimal performance in our silly laboratory task[s]
>
> Anderson (1990, p. 42)

The idea that we will fail to grasp what memory *is* without knowing what memory *is for* is hardly new. Several authors have presented this idea before (Bjork & Bjork, 1992; Neisser, 1978) others have developed specific (McClelland & Chappell, 1998; Shiffrin & Steyvers, 1997) or even quite general models (Anderson, 1990) around it. In this paper, we join these authors by proposing a narrow version of the same idea. We propose that we encode and store information as a function of the particular ways we have used similar information in the past. More specifically, we contend that the experience of retrieval can serve as a powerful cue to the most effective ways to encode similar information in comparable future learning episodes.

In order to explore these notions, we designed two experiments in which all participants went through multiple study–test cycles of single-category lists with different types of tests. Each

Address correspondence to: Leonel Garcia-Marques, Faculdade de Psicologia, Alameda da Universidade, 1648-013 Lisboa, Portugal. E-mail: garcia_marques@sapo.pt
Preparation of this manuscript was supported in part by Grant [POCTI/PSI/66864/2006] from Fundação para a Ciência e a Tecnologia (FCT—Portugal).

test involved recognition, but either included only same-category lures or only different-category lures. From here onwards, we refer to this manipulated feature as "test structure". We expected that the experience of multiple study–test cycles (i.e., repeated testing) would lead to adaptive changes in subsequent performance as long as test structure remained constant; but this adaptive change would also carry a cost that would be detectable whenever the test structure changed.

As our goals and methods combine ideas that have been explored under the rubrics of test expectancy (e.g., Balota & Neely, 1980; Finley & Benjamin, 2012), the power of testing (e.g., Landauer & Bjork, 1978; Roediger & Karpicke, 2006) and conceptual learning effects (Higham & Brooks, 1997; Kantner & Lindsay, 2010), we will first briefly refer to these literatures, both to review previous findings on the effects of test experiences on subsequent encoding, and to describe how our studies differ from the studies carried out under these rubrics.

Testing expectancy

The first studies identifying test expectancy effects were reported by Meyer (1934, 1936). Meyer's findings were intuitive and compelling: students expecting an essay test always performed better than those expecting a multiple-choice test, regardless of final criterial test format. The take home message appeared to be clear—students should always study as if for an essay test, even when they expect a different kind of test. Further lab research used practice tests to induce expectancy of a final criterial tests, and roughly equated essays with free recall and multiple-choice with recognition. Findings from these studies either converged with the conclusion that expectations of a recall test produced superior performance in all cases (e.g., Balota & Neely, 1980; Hall, Grossman, & Elwood, 1976, exp. 1; Neely & Balota, 1981; Schmidt, 1983; Thiede, 1996) although some studies only found the recall expectancy superiority in recall performance (e.g., Connor, 1977; d'Ydewalle, 1981; Hall et al., 1976, exp. 3; Maisto, DeWaard, & Miller, 1977). Puzzling enough, the picture is even less clear-cut in field studies, where a test expectancy match benefit is more often found than recall expectancy superiority effects—students expecting an essay do better at essay tests than students expecting multiple-choice test; whereas the converse is true for multiple-choice performance (for a meta-analysis that includes both lab and field studies, see Lundeberg & Fox, 1991).[1] With some exceptions (e.g., Finley & Benjamin, 2012; Leonard & Whitten, 1983; Tversky, 1973; Whitten & Leonard, 1980), the test expectancy literature has been targeted at exploring and accounting for Meyer's initial findings, either in lab studies that induced recall or recognition expectancies by practice tests or in field studies that used the supposedly more naturalistic test counterparts, namely essay and multiple-choice tests and induced test expectancy by manipulating practice tests, instructions or both. In our case, although we use practice tests to induce test expectancies just as many of the lab studies in this area, we are not interested in how test expectancies affect test performance per se, instead we are more interested in exploring how experience with testing leads to more adaptive encoding of subsequent information for the same type of test. More specifically, we are interested in showing how test expectancy and beliefs about test structure can influence future study episodes and not only performance on tests.

The power of testing

Since its inception (Ebbinghaus, 1885), the scientific study of memory has maintained a sharp contrast between the phases of study and test. Learning would—ideally—occur only during the study phase and testing would serve exclusively to provide a more or less sensitive assessment of learning. It is true that over time there have been a few critical voices speaking up against this simplistic view (e.g., Bjork, 1975; Izawa, 1971), but by and large in the memory literature, the investigation of encoding has been separated from the study of retrieval. Moreover, research on encoding processes has predominated over research on retrieval processes. Recently, under the concept of the "testing effect" (for recent reviews, see Roediger & Butler, 2011; Roediger, Putnam, & Smith, 2011), the literature has uncovered a host of processes that occur at testing and that can modify, add and generally boost learning. For instance, Karpicke and Roediger (2008) have shown that repeated retrieval practice (in the absence of new study opportunities) enhances

[1]There were however a few exceptions, even for the more restricted version of Meyer's hypothesis (e.g., McDaniel, Blischak, & Challis, 1994)

long-term learning much more than repeated studying of the same information (in the absence of repeated testing).

Our studies, however, are not concerned with the benefits of repeated testing of the same information and long-term retention, but with the changes induced by repeated testing in the encoding of new information. That is, we are interested in test potentiation effects, a much lesser explored subject. Research on test potentiation is usually related to how unsuccessful retrieval attempts enhance learning by improving encoding of the same or related items in subsequent study episodes (Finley & Benjamin, 2012; Izawa, 1971; Karpicke, 2009; Szpunar, McDermott, & Roediger, 2007; Wissman, Rawson, & Pyc, 2011). Our studies depart from this more usual line of research in the sense that we aim to study how the development of strategies during retrieval can change encoding of a different set of items. Despite this difference, we are still interested in a mediated testing benefit and not just in the direct benefits of testing.

Mediated and direct effects of testing have often been treated in an undifferentiated way and only recently have testing and test potentiation effects been experimentally isolated (see Arnold & McDermott, 2013) and the possible processes by which potentiation occurs clarified (Grimaldi & Karpicke, 2012). Our aim with this study is precisely to enlighten some of the mechanisms that might be responsible for some forms of the test potentiation effect, namely with novel items.

Conceptual learning

In our studies, although all study lists correspond to exemplars of a given category, in one condition, the recognition test includes only lures from a different category. In this sense, our recognition test becomes very similar to classification tasks used in concept learning (Bruner, Goodnow, & Austin, 1956; for an extensive review, see Murphy, 2002; for a review on implicit learning, see Seger, 1994). In classification tasks, participants have to learn to discriminate items pertaining to different categories (corrective feedback is usually provided during discrimination). Performance is usually assessed by classification accuracy of new exemplars. In recognition memory tasks participants have to discriminate items pertaining to the study list from lures (new items), and corrective feedback is usually not provided during discrimination. Performance is usually assessed by identification accuracy ($P_{(\text{“old”/“old”})} - P_{(\text{“old”/“new”})}$). But classification and recognition tasks can be very similar because participants in classification tasks are sometimes asked to study examples of category A, and only later are they asked to classify old and new items members as members of category A or B (non-A). In classification tasks, the discrimination criteria are feature based, corrective feedback is usually provided, and the learning episodes are not clearly separated from the testing episodes as usually occurs in memory tasks. In recognition memory tasks, the discrimination criteria are episodic, special care is taken to prevent feature-based learning (i.e., list items should match foils in every respect other than their encoding status), corrective feedback is not usually provided, and the separation between learning and testing phases is clear-cut, with the tasks for each phase being clearly different. These differences conspire to allow for little concept or feature-based learning to occur, or at least to be ascertainable in recognition tasks. This is unfortunate, because memory and concept-based learning are likely to be closely interwoven in the real world. In fact, the literature has already documented a host of effects of knowledge on memory (e.g., prototype memory effects and false memory effects based on previous associations or knowledge, see Anisfeld & Knapp, 1968; Brainerd & Reyna, 1988; Posner & Keele, 1968; Roediger & McDermott, 1995; Underwood & Freund, 1968) and several theoretical approaches have already attempted to delineate a more meaningful relationship between memory and concept learning (Brainerd & Reyna, 2001; Hintzman, 1986). More specifically, Higham and Brooks (1997) have obtained classical recognition memory effects in a classification task and classical classification effects in a recognition memory task. Thus, in a classification task, participants classified items as better category exemplars when they had a positive episodic status (i.e., they were “old”) or when they were deeper processed (see Craik & Lockhart, 1972). Moreover, these results reversed for new items such that a “mirror effect” was obtained (see Glanzer & Adams, 1985). On the other hand, in recognition tests, these authors obtained common classification effects, such as centrality effects—false recognition of new exemplars increases whenever their features are congruent with the features of old items. Given these findings, we believe that inducing test expectancy by retrieval practice will allow us to uncover systematic concept learning effects in recognition and to further characterise the relationship between concept learning and memory.

EXPERIMENT 1

In Experiment 1, participants went through four study–test cycles. Each study list was formed from exemplars of a common category, with a different category used for each study–test cycle. For the two between-subjects conditions, the first three cycles differed only in the nature of the recognition test. In the Related Lure condition, all lures were exemplars of the category presented in the study list, matched to the presented items in terms of frequency of production—i.e., mean number of times the item is produced in a category generation task. In the Unrelated Lure condition, all lures were exemplars of a different category (one category per list) from the one presented in the study list. In the last study–test cycle, participants in both conditions received a test like the ones participants had received in the Related Lure condition (i.e., with all lures being exemplars of the same category presented in the study list). Our goal was to examine how participants adapt their encoding strategies to the requirements of the situation in which the learned information is to be retrieved. In this regard, the two groups of participants stood in stark contrast. For the first three study–test cycles, for participants in the Unrelated Lure condition, the identification of the category of the exemplars included in each study list was crucial and sufficient to assure good performance in the subsequent recognition test; whereas for participants in the Related Lure condition, identifying the category to which all of the items belonged was of little value. More specifically, in the Related Lure condition, category knowledge had no diagnostic power when applied to the recognition test because the lures were matched to the presented items in terms of category membership and frequency of production (i.e., centrality). Thus, participants in the Unrelated Lure condition should make their recognition judgments based on the category membership of test items, whereas participants in the Related Lure condition should base their recognition judgments on the distinctive features of study list members. So, we expect that in test 4, participants in the Unrelated Lure condition fail to adapt to the new test and show more false alarms and possibly less hits than participants in the Related Lure condition, who should already have adapted their encoding strategy to the type of test.

Although we are mostly interested in the performance on test 4, we are confident that recognition performance on the first three tests will be affected by our test structure manipulation, and that participants in the Unrelated Lure condition will greatly outperform participants in the Related Lure condition on these three tests. Also, a more interesting question about performance on the three first tests arises—will the difference in the types of lures included in the recognition test affect criterion in a contrasting way? More specifically, participants in the Unrelated Lure condition should exhibit a highly restrictive criterion in that they should reject all the lures (which are from a different category than the study list items) as well as the least typical studied items because low centrality items are often not recognised as category members (Greenberg & Bjorklund, 1981; Loftus, 1975; McCloskey & Glucksberg, 1978). On the other hand, participants in the Related Lure condition should exhibit the opposite response tendency (i.e., a lenient criterion) because all items on the test are from the study list item category. This leniency bias should be greater for high centrality relative to low centrality items. It is important to note that the stability of the criterion used across the study–test cycles should differ between conditions. Participants in the Related Lure condition should gradually learn to avoid using conceptual knowledge to inform their recognition judgments. Thus, their criterion should become less lenient across the four study–test cycles and the centrality bias should also decrease across cycles. On the contrary, participants in the Unrelated Lure condition should maintain a stable criterion until the fourth test, which involves a sudden change in the nature of the test (where all lures belong to the study list item category). This change in the nature of the lures should lead to a dramatic change in criterion for participants who had experienced only unrelated lures on previous tests.

Method

Participants. Eighty-eight Lisbon University undergraduates participated in the experiment for payment. Forty-nine participants were assigned to the Unrelated Lure condition (tested with unrelated lures in tests 1–3) and 39 were assigned to the Related Lure condition (tested with related lures in tests 1–3).

Materials. Four categories were selected from the Portuguese category norms (Pinto, 1992)—animals, fruits, occupations and body parts—to construct the categorical lists and the related lures. These four categories were chosen because they were the categories with the greatest number of exemplars with a category frequency (i.e., probability of being generated as a category exemplar) above 5%. For each category, the 45 most frequent exemplars were selected. Exemplars with more than one word were excluded.

For each category, three study lists were built so that, when ordered in terms of decreasing frequency, items in positions 1 and 4, 2 and 5, or 3 and 6 were not presented and could be used as lures in the recognition tests for participants in the Related Lure condition. As list words and lures were counterbalanced between participants, three different study lists per category were created, and each participant studied one of these lists per category. Each study list was composed of 30 exemplars of the same category randomised afresh for each participant. For each study list, two types of recognition tests were created. Both types of test included 15 out of the 30 studied words. But the Related Lure recognition test included the 15 unstudied category members of the studied category, whereas Unrelated Lure recognition test included 15 unstudied and unrelated words, selected from 15 other non-presented categories. So, in both conditions, each recognition test included a total of 30 words. The unrelated lures used in the three recognition tests of the Unrelated Lure condition were drawn from a set of 45 exemplars of 15 non-presented categories. For each recognition test this set was divided into three sets, so that 15 lures of 15 different categories were included. Thus, for the Unrelated Lure condition, the lures were different for each one of the three tests, and each lures on a single test came from a different category—but the 15 categories were repeated across the three tests. Test 4 was identical for the two conditions and was composed of the 15 list words and the 15 unstudied members of the same category.

Design. The design was between participants, so participants in the Related Lure condition received tests with lures that were from the same category as studied words in tests 1–3, whereas participants in the Unrelated Lure condition received tests with lures that were from different categories in tests 1–3. Test 4 was identical for all the participants, and included related lures, varying only in terms of category and the specific lures that were used for counterbalancing purposes. The study lists were counterbalanced using a Latin square design, resulting in four different study list orders. The level of correct recognition and false alarms was measured for all tests, so the criterion and sensibility could be calculated for all tests, as well as performance differences (in this case, the comparison of interest was on test 4).

Procedure. Participants were told that they were going to study lists of words, and that their memory about those words would be tested afterwards. Words were presented one by one in the centre of a computer monitor at the rate of 2000 ms per word. After each list (30 words) presentation, all participants solved math problems for 1 minute. After this distractor task, participants received classic recognition instructions, and those in the Related Lure condition received a recognition test composed of 15 presented words and the 15 unpresented category members, while participants in the Unrelated Lure condition received a recognition test composed of 15 presented words and 15 unpresented unrelated words. This cycle of study-math-test was repeated for the four lists. Test 4 was identical for participants in the two conditions, being composed of 15 presented words and 15 related distractors, i.e., same category members. For the recognition tests, each word was presented one by one in the centre of the computer monitor, and word order was randomised for each participant. The recognition tests were self-paced and participants responded by using the keyboard. Participants were instructed to press the key "c" if they thought the word had been presented before or the key "m" if they thought the word had not been presented (the key "m" had a red sticker on it and the key "c" had a green sticker on it, and participants were told that they should press the red key if they had not seen the word before and the green key if they had seen the word before).

Results and discussion

We first describe the results in terms of hit and false alarm data, and then turn to discrimination and criterion measures, which were statistically analysed to address our hypotheses. Mean hits and

TABLE 1
Mean hits and false alarms from the four study–test cycles in the Related and Unrelated Lure conditions for Experiment 1

	Related lures				*Unrelated lures*			
	Cycle 1	*Cycle 2*	*Cycle 3*	*Cycle 4*	*Cycle 1*	*Cycle 2*	*Cycle 3*	*Cycle 4*
Hits	.87 (.10)	.84 (.09)	.81 (.13)	.76 (.17)	.87 (.11)	.90 (.07)	.88 (.10)	.83 (.13)
FAs	.30 (.16)	.27 (.20)	.31 (.22)	.29 (.19)	.06 (.06)	.06 (.03)	.06 (.03)	.41 (.28)

Note: Cycle 4 for Unrelated Lures condition refers to a recognition test containing related lures only. Standard deviations are in parentheses.
FAs, false alarms.

false alarms from the four study–test cycles in the Related and Unrelated Lure conditions are presented in Table 1. The data show a mirror effect. That is, participants in the Unrelated Lure condition produced both more hits and fewer false alarms than participants in the Related Lure condition on tests 1 – 3. On the fourth test, which was the same in the two conditions, participants in the Unrelated Lures condition produced more hits as on previous lists, but also more false alarms.

To explore the effect of the lure manipulation on performance throughout the study–test cycles, we subjected both d′ (ability to discriminate between studied items and lures) and c (criterion or response bias; reflects the degree to which "yes"/"old" responses dominate or the degree to which "no"/"new" responses are preferred; negative values indicate a bias towards classifying items as "old", and positive values indicate a bias towards classifying items as "new"; Macmillan & Creelman, 2005) to a 2 Lure (Related vs. Unrelated) × 3 Study–Test Cycle (1–3) mixed-model analysis of variance (ANOVA) and then we separately test the results of our manipulation in the fourth study cycle (in which the participants of both conditions received the same related lure recognition test). As a reminder, note that Lure was manipulated between subjects. Let us first present the results for the memory discrimination measure, d′. We obtained a main effect of Lure, showing that participants in the Unrelated Lure condition (M = 2.89) outperformed participants in the Related Lure condition (M = 1.73), $F(1, 86)$ = 149.43, $p < .001$, MSE = .57, η^2 = .63. We also obtained a marginal main effect of study–test cycle, $F(2, 86)$ = 2.90, p = .056, MSE = .26, η^2 = .03, but most importantly, the two factors interacted showing that the superior performance of participants in the Unrelated Lure condition relative to participants in the Related Lure condition disappears on the fourth test, $F(3, 86)$ = 29.68, $p < .001$, MSE = .32, η^2 = .18. These results were expected since the recognition is much easier for Unrelated Lure participants. In the fourth study cycle (in which the participants of both conditions received the same related lure recognition test), however, the two conditions no longer differed ($t < 1$). Figure 1 depicts all the relevant d′ scores data.

In the analysis of c (criterion or response bias), we also obtained a main effect of lure, such that participants from the Unrelated Lure condition (M = .15) showed a more biased criterion than participants in the Related Lure condition, who were more liberal (M = –.23), $F(1, 86)$ = 48.65, p = .001, MSE = .19, η^2 = .36. We also obtained the interaction showing that the two factors interacted, $F(2, 86)$ = 3.27, p = .040, MSE = .06, η^2 = .04. To interpret this interaction we contrasted the linear trend between the two conditions and it was indeed different, $F(1, 86)$ = 3.78, p = .055, MSE = .05. And this happened because, whereas Unrelated Lure participants did linearly decrease their bias across study–test cycles, $F(1, 86)$ = 4.33, p = .040, MSE = .05, the same did not occur for

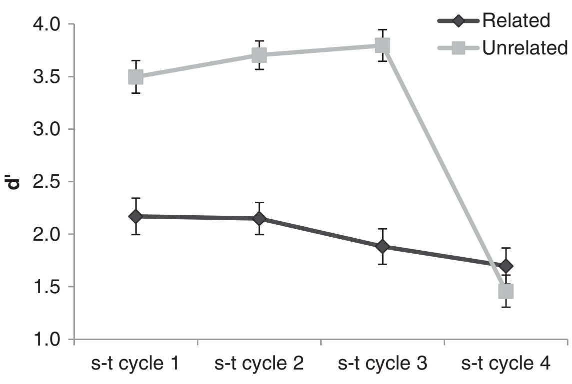

Figure 1. Study–test × lures condition interaction on d′ scores in Experiment 1. Vertical bars denote standard errors of the means.
Note: S-T cycle 4 for Unrelated Lures condition refers to a recognition test containing related lures only.

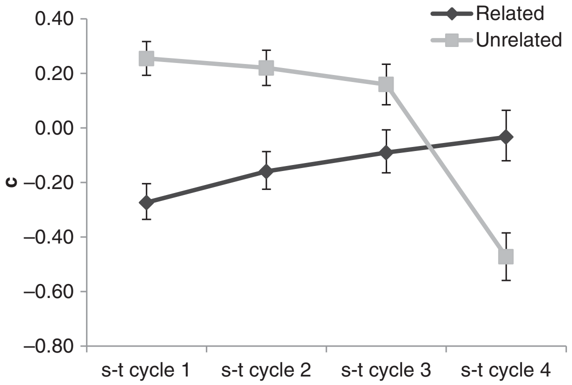

Figure 2. Study–test cycle × lures condition interaction on bias (c) in Experiment 1. Vertical bars denote standard errors of the means.
Note: S-T cycle 4 for Unrelated Lures condition refers to a recognition test containing related lures only.

Unrelated Lure participants ($F < 1$). Finally, in the fourth trial (in which the participants of both conditions received the same related lure recognition test), the initial differences between Unrelated and Related Lures participants reversed such that the bias was now stronger and more liberal for Unrelated Lure ($M = -.40$) relative to Related Lure $M = -.08$) participants, with a $F(86) = 7.08$, $p = .001$, MSE = .33, $\eta^2 = .08$. Figure 2 depicts all the relevant c scores data.

To further explore the role of conceptual knowledge on criterion placement we performed a median split by centrality of the items and lures used in the Related Lures condition. We had to restrict this analysis to the Related Lures condition because in the Unrelated Lures condition, lures were taken from a different category than the studied items (except for the fourth study cycle). We present hits and false alarm data in Table 2. Data from Table 2 show a centrality effect such that high centrality targets and lures are more readily accepted as old items than the corresponding low centrality items and lures. To analyse these data, we ran two centrality (High vs. Low) × 4 Study–Test Cycle within-subjects ANOVAs on both d′ and c. The results of the ANOVA for the memory discrimination measure, d′, only produced a centrality main effect, showing that participants' performance was superior for low centrality items and lures ($M = 2.36$) relative to high centrality items and lures ($M = 2.13$), $F(1, 38) = 488$, $p = .033$, MSE = .68, $\eta^2 = .11$.

Figure 3 shows criterion in the Related Lure condition by centrality and study–test cycle. In the analysis relative to criterion measure c, we also obtained a main effect of centrality, such that participants more readily accepted high centrality items and lures ($M = -.30$) relative to low centrality items and lures ($M = -.05$), $F(1, 38) = 32.92$, $p < .001$, MSE = .15, $\eta^2 = .46$. We also obtained a main effect of study–test cycle, $F(3, 114) = 2.66$, $p = .051$, MSE = .39, $\eta^2 = .06$, mainly due to the fact that the overall criterion becomes dramatically less and less biased across the study–test cycles. And indeed the linear trend is significant, $F(1, 38) = 7.32$, $p = .010$, MSE = .41

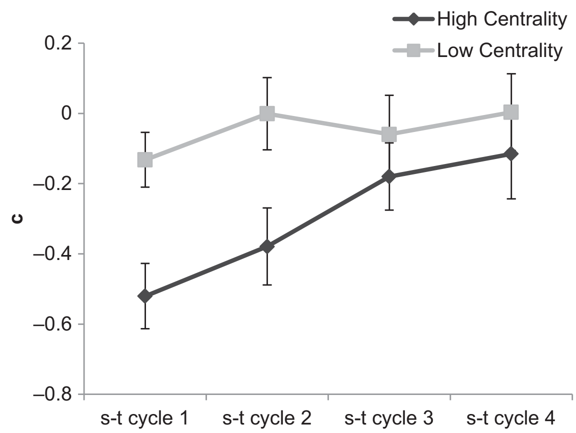

Figure 3. Changes in bias (c) across the four study–test cycles for high and low centrality words (Related Lures condition only) in Experiment 1. Vertical bars denote standard errors of the means.

TABLE 2
Mean hits and false alarms for high and low centrality items and lures (Related Lures condition only) along the study–test cycles for Experiment 1

	High centrality				*Low centrality*			
	Cycle 1	*Cycle 2*	*Cycle 3*	*Cycle 4*	*Cycle 1*	*Cycle 2*	*Cycle 3*	*Cycle 4*
Hits	.90 (.13)	.86 (.12)	.82 (.17)	.77 (.21)	.86 (.14)	.84 (.12)	.80 (.17)	.76 (.20)
FAs	.33 (.25)	.34 (.30)	.27 (.24)	.28 (.25)	.20 (.19)	.18 (.19)	.23 (.22)	.25 (.23)

Note: Standard deviations are in parentheses.
FAs, false alarms.

and the effect size is considerably larger ($\eta^2 = .16$) than the overall main effect whereas the residuals were non-significant, $F < 1$. Critically, although the interaction was not significant, $F(3, 114) = 2.99$, $p = .119$, MSE = .23, $\eta^2 = .05$, the above-mentioned linear trend was indeed moderated by centrality, $F(1, 38) = 6.65$, $p = .015$, MSE = .17, $\eta^2 = .15$, whereas the residuals were non-significant, $F < 1$.

Finally, we also compared contrasted high and low centrality across Unrelated and Related Lure conditions in the fourth cycle (in which the participants of both conditions received the same related lure recognition test) for both d′ and c. Thus we first performed a 2 Lure (Unrelated vs. Related) × 2 Centrality (High vs. Low) mixed ANOVA on d′ scores, with the last factor being within participants. We obtained a main effect for centrality, $F(1, 86) = 3.76$, $p = .056$, MSE = 1.05, $\eta^2 = .04$, such that participants performed better with Low ($M = 1.88$) than with high centrality ($M = 1.56$) items and lures. A marginal interaction also emerged, $F(1, 86) = 2.92$, $p = .091$, MSE = 1.05, $\eta^2 = .03$, suggesting that the differences between low and high centrality items occurred more pronouncedly in the Unrelated ($M = 1.82$ vs. $M = 1.25$) than in the Related Lure ($M = 1.94$ vs. $M = 1.90$) condition. As for the same analysis with c scores, we only replicated the already reported difference between conditions in the last study cycle and this difference was not qualified by centrality ($F < 1$).

We interpreted these data as evidence that, across the study–test cycles, participants in both conditions tried to adapt their encoding and test strategies to the requirements of the specific test they had been faced with. Thus, whereas participants in both conditions probably identified the taxonomic category underlying the first study list, this conceptual knowledge facilitated the performance of participants in the Unrelated Lure condition but it hindered performance of participants in the Related Lure condition, because the lures were from a different category in the former condition but from the same category in the latter condition. More specifically, participants' conceptual knowledge leads to a liberal criterion in the Related Lure condition (especially in the case of high centrality) because both targets and lures belonged to the same category; and to a stringent criterion setting in the case of the Unrelated Lure condition because all lures belonged to a different category, which made lure rejection very easy (along with a certain degree of rejection of low centrality items). Across the study–test cycles, participants in the Related Lure condition learned to disregard their knowledge about the category underlying the study lists and their criterion became less and less biased (lenient) overall and in especially for high centrality targets and lures. However, the same did not occur in terms of memory discrimination. Apparently, the constant change of categories across the study–test cycles may have limited the possibility of developing a strategy that would improve memory performance, while at the same time may have facilitated insight regarding the low diagnosticity of category membership.

Participants in the Unrelated Lure condition, on the contrary, kept relying on their conceptual knowledge to perform on the recognition test, such that when on the last study–test cycle the structure of the test was changed to include only lures from the same category as the studied items, participants' criterion placement showed a greater degree of bias than the initial bias shown by the participants in the Related condition.

Thus, participants in both conditions initially relied on their conceptual knowledge, showing the centrality effect, in other words being more likely to endorse high centrality items as old. However, across the first three study–test cycles, participants in the Related Lure condition learned to avoid this conceptually based response because it promoted limited performance. The same did not happen to participants in the Unrelated Lure condition, for whom the conceptually based response strategy allowed them to achieve very high levels of performance until the last critical study–test cycle. Thus, these participants in the Unrelated Lure condition were never given the opportunity to learn how to avoid this response strategy and when they were confronted with a recognition test in which all the lures were from the same category as the studied items (the fourth test), they exhibited an even more liberal response criterion than they would probably have shown in the absence of previous tests (i.e., a criterion probably similar to the criterion shown by participants in the Related condition when they were tested for the first study–test cycle).

EXPERIMENT 2

We contend that the requirements of experienced tests shape the way participants encode similar new information or new information in similar contexts. These test-driven differences of encoding should affect performance on similar tests but they can also affect performance on different tests if participants are surprised by an unwarned change in test format. In Experiment 2, the new test we use is free recall. We relied on the same Related/Unrelated Lure manipulation as in Experiment 1, but on the fourth and last test we asked our participants to perform a free recall test, and we did not warn them of this change in test format. We reasoned that participants in the Related Lure condition would learn a strategy based on the individual features of the studied items (a distinctive feature-based strategy), since a category-based strategy would be quite inefficient given the structure of the test lists they had experienced, whereas participants in the Unrelated Lure condition would learn to rely precisely on the categorical or thematic nature of studied items (a relational, conceptual-based strategy). Since feature-based encoding strategies enhance free recall of lists of related items (Einstein & Hunt, 1980; see also, Hunt & McDaniel, 1993; Nairne, 2006) we expect that despite the Unrelated Lure participants outperforming the Related Lure participants across the first three study–test cycle, this performance difference would be reversed in the final study–test cycle, in which the test requires free recall.

Method

Participants. Ninety-four Lisbon University undergraduates participated in this experiment for payment. Forty-six participants were assigned to the unrelated condition (tested with unrelated distractors in tests 1–3) and forty-eight were assigned to the related condition (tested with related distractors in tests 1–3).

Materials, design and procedure. The same materials, design and procedure used in Experiment 1 were also used in this Experiment, with two exceptions: first, the final test was a free recall test rather than a recognition test—after studying the fourth list and solving math problems, the participants were instructed to write on a blank sheet as many words as they could recall from the last list of words; and second, in the first three recognition tests the participants' responses were given on a sheet of paper, where all the test items were listed, rather than on the computer. In the study phase, the words were presented one by one on the computer screen (randomised afresh for each participant), but after each list presentation, followed by the math distractor task, the participants were instructed to read the words written on the sheet of paper and to decide whether each word had been presented on the previous list. For each word, participants were asked to circle "yes" if they thought that the word had been presented and "no" if they thought it had not. The recognition test was administered in a self-paced manner and the test words were presented in a random order for each version (without the items from positions 1 and 4, 2 and 5 and 3 and 6) and list presentation order (as determined by a Latin square design), resulting in a total of 12 test word presentation orders for each condition. The randomization of the words on each test was constant because the tests were administered on paper.

Results and discussion

Hits and false alarms from the first three study–test cycles in both conditions are presented in Table 3. The data show that participants in the Unrelated Lure condition produced about the same level of hits but noticeably fewer false alarms relative to participants in the Related Lure condition. To explore the effect of the lure manipulation on recognition performance throughout the first three study–test cycles, we performed 2 Lure (Related vs. Unrelated) × 3 Study–Test Cycle (1–3) mixed-model ANOVAs on both d′ and c. Regarding the

TABLE 3
Mean hits and false alarms from the three study–test cycles in the Related and Unrelated Lure conditions for Experiment 2

	Related lures			*Unrelated lures*		
	Cycle 1	*Cycle 2*	*Cycle 3*	*Cycle 1*	*Cycle 2*	*Cycle 3*
Hits	0.89 (.10)	0.84 (.14)	0.84 (.15)	0.84 (.18)	0.88 (.17)	0.84 (.09)
FAs	0.20 (.13)	0.20 (.15)	0.18 (.13)	0.05 (.13)	0.02 (.11)	0.01 (.05)

Note: Standard deviations are in parentheses.
FAs, false alarms.

memory discrimination measure, d′, we only obtained a main effect of Lure, showing that participants in the Unrelated Lure condition (M = 3.51) outperformed participants in the Related condition (M = 2.45), $F(1, 90) = 64.89$, $p < .001$, MSE = 1.70, $\eta^2 = .42$. In the analysis relative on the criterion measure, c, we again obtained a main effect of Lure, such that participants in the Unrelated Lure condition (M = .45) showed a more stringent criterion than participants in the Related Lure condition (M = −.14), $F(1, 90) = 82.63$, $p < .001$, MSE = .29, $\eta^2 = .48$. We also obtained a main effect of study–test cycle, $F(3, 90) = 4.48$, $p = .013$, MSE = .09, $\eta^2 = .05$, mainly due to the fact that the overall criterion became more stringent across the three study–test cycles. The two factors did not interact significantly, $F(2, 180) = 2.37$, $p = .100$, MSE = .09, $\eta^2 = .02$. Thus, the difference in bias between Lure conditions remained reliable across the three study–test cycles. However, across the three study–test cycles, participants in the Related Lure condition exhibited a positive linear trend, $F(1, 90) = 5.99$, $p = .016$, MSE = .12, $\eta^2 = .06$ (with a non-significant residuals, $F < 1$), whereas this linear trend was absent from the data of participants in the Unrelated Lure condition, $F(1, 90) < 1$. Figure 4 depicts these data.

Again, we conclude that across the first three similar study–test cycles, participants in both conditions tried to adapt their encoding and test strategies to the requirements they were confronted with. More specifically, as in Experiment 1, participants' conceptual knowledge lead to a liberal criterion setting for participants in the Related condition and to a stringent criterion setting for participants in the Unrelated Lure condition. Across the study–test cycles, participants in the Related Lure learned to disregard their knowledge about the categories underlying the study

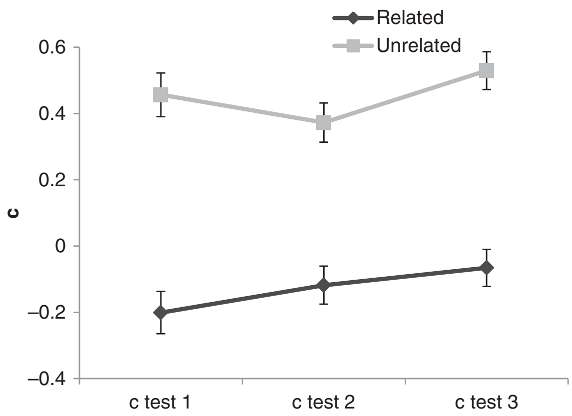

Figure 4. Changes in bias (c) for the first three study–test cycles for Related and Unrelated Lures participants in Experiment 2. Vertical bars represent standard errors of the means.

lists and their criterion became less liberal. However, and like in Experiment 1, this did not occur in terms of memory discrimination. Participants in the Unrelated Lure condition, on the contrary, maintained their encoding and response strategy based on conceptual knowledge to perform in the recognition test.

Like in Experiment 1, to further explore the role of conceptual knowledge on criterion placement, we split both items and lures used in the Related Lure condition into low and high centrality lures. We present the hits and false alarm data for these two types of items in Table 4. Data from Table 4 shows a centrality effect such that high centrality lures are more readily falsely recognised than low centrality lures. To follow-up in the analyses of these data, we computed two 2 Centrality (High vs. Low) × 3 Study–Test Cycle mixed-model ANOVAs (the last factor being within-participants) on both d′ and c. Two main effects emerged from the results for the memory discrimination measure, d′. Thus, a centrality main

TABLE 4
Mean hits and false alarms for high and low centrality items and lures (Related Lures condition only) along the first three study–test cycles for Experiment 2

	High centrality			*Low centrality*		
	Cycle 1	*Cycle 2*	*Cycle 3*	*Cycle 1*	*Cycle 2*	*Cycle 3*
Hits	0.88 (.19)	0.84 (.16)	0.81 (.18)	0.86 (.19)	0.86 (.14)	0.87 (.15)
FAs	0.27 (.23)	0.23 (.18)	0.24 (.19)	0.16 (.18)	0.17 (.18)	0.13 (.14)

Note: Standard deviations are in parentheses.
FAs, false alarms.

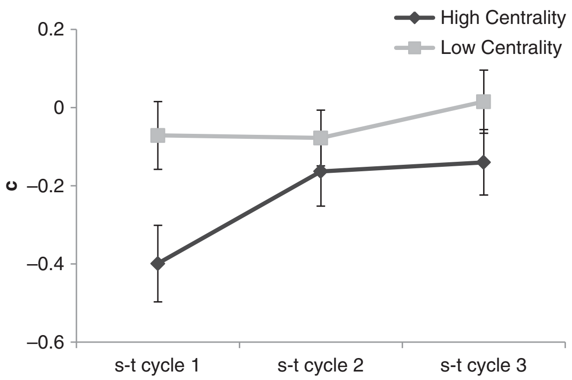

Figure 5. Changes in bias (c) across the first three study–test cycles for high and low centrality words (Related Lures condition only) in Experiment 2. Vertical bars denote standard errors of the means.

effect emerged, indicating that participants' performance was superior for low centrality items and lures (M = 2.80) relative to high centrality items and lures (M = 2.29), $F(1, 46) = 20.28$, $p < .000$, MSE = .91, $\eta^2 = .31$. A study–test cycle's main effect was also significant, due to the better performance on the first study–test (M = 2.76) relative to the remaining cycles (M = 2.39 and M = 2.48), $F(2, 92) = 3.25$, $p = .043$, MSE = 1.09, $\eta^2 = .07$.

In the analysis relative to criterion measure c, we also obtained a main effect of centrality. The main effect of centrality indicated that participants more readily accepted high centrality targets and lures (M = −.23) relative to low centrality targets and lures (M = −.04), $F(1, 46) = 6.50$, $p < .014$, MSE = .39, $\eta^2 = .12$. Critically, although the interaction was not significant, $F(2, 92) = 1.41$, $p = .249$, MSE = .26, $\eta^2 = .03$, a difference in the acceptance of high centrality targets and lures (M = −.40) and low centrality targets and lures (M =.−.07) emerged at the first study–test cycle, $F(1, 46) = 6.89$, $p = .012$, MSE = .37, $\eta^2 = .13$, but failed short from significance for the remaining cycles (both Fs < 1). Figure 5 depicts these data.

We interpret these data as replicating results from Experiment 1, in that although participants' conceptual knowledge leads to a liberal criterion setting in the Related Lure condition, across the study–test cycles, participants learned to disregard their knowledge about the category underlying the study lists and their criterion became less biased. This effect could be documented in both the overall gradual "de-liberalising" of the criterion and in the gradual reduction of difference in criterion between high and low centrality items across the study–test cycles.

Finally, we compared free recall in the fourth study–test cycle between the Related and Unrelated Lure conditions. The number of intrusions was small and not significantly different between the two conditions, so we ignored them in subsequent analyses and just focused on correct recall. Performance differed by Lure condition, $t(90) = 2.70$, $p = .001$, SD = 5.60, $d = .56$, such that participants in the Related Lure condition (M = 13.55) outperformed participants in the Unrelated Lure condition (M = 10.40). To check whether the centrality effect found across the previous study–test cycles in recognition would also be obtained in free recall, we split the recalled items in high and low centrality and computed a 2 Lure condition (Related vs. Unrelated) × 2 Centrality (High vs. Low) mixed-model ANOVA, the last factor being within participants. The analysis revealed two main effects, a Lure main effect that of course reproduced the previously obtained difference, and a centrality main effect, such that high centrality targets (M = 6.70) were better recalled than low centrality targets (M = 5.50), $F(1, 91) = 14.89$, $p < .001$, MSE = .22, $\eta^2 = .14$. The interaction was not significant, $F < 1$. Thus, although participants in the Related condition learned how to overcome a knowledge-based bias in their responses to recognition tests, such that the centrality effect as measured by criterion was gradually overcome, a centrality effect reappeared for these participants when the retrieval context was no longer detrimental to such a response strategy (i.e., at the free recall test).

As we have argued that test requirements lead to an adaptive change in encoding and test strategies across the study–test cycles, and that it was these strategy changes that accounted for the differences obtained in the last study–test cycle (in the free recall test), we measured the correlation between d′/c and the number of items recalled. Because test requirements varied as a function of condition, we computed these correlations separately for the Related and Unrelated Lure conditions. These correlations are depicted in Table 5.

Data from Table 5 suggest that, for participants in the Related Lure condition, the encoding and response strategies that allowed them to discriminate between list items and lures in the first three study–test cycles were similar to the strategy they adopted in the last study–test cycle, as the correlations between d′ scores across the study–test and the number of recalled items were all statistically significant and strong. No correlation was found between c scores and the number of items

TABLE 5
Correlations between the d′ and c scores obtained in the first three study–test cycles and the number of items recalled in the final cycle (computed by Lures condition) in Experiment 2

	d′			c		
	Cycle 1	Cycle 2	Cycle 3	Cycle 1	Cycle 2	Cycle 3
Free recall (*Related lures*)	.61*	.59*	.51*	−0.09	−0.03	0.04
Free recall (*Unrelated lures*)	0.13	0.27	0.17	−0.03	−0.17	−.31*

*$p < .05$.

recalled. On the contrary, participants in the Unrelated Lure condition showed a completely different performance pattern. The encoding and response strategies that led them to successfully discriminate between items and different-category lures seem to be quite different from the strategy these participants used in recall. This is, for participants in the Unrelated Lure condition, the correlations obtained between d′ scores and the number of items recalled were small and non-significant. On the other hand, for these participants, response criterion on the last recognition test did predict the number of items recalled on the free recall test, such that the more stringent the criterion, the fewer items were recalled in the last study–test cycle (free recall test). As we interpreted this stringency in criterion as evidence of knowledge-based encoding and response strategy, this result concurs with our interpretation.

GENERAL DISCUSSION

In this paper, we argued that specific requirements of retrieval contexts in which previous learning is assessed affect future encoding of similar information or information acquired in similar contexts. In two studies, in which the instructional sets were always identical across conditions (i.e., memorise categorised study lists), we simply altered the nature of the recognition tests by including related (same-category) or unrelated (different-category) lures at test. This difference guaranteed a higher performance from participants in the Unrelated Lure condition and promoted the setting of a conceptual-based stringent response criterion that remained stable until the retrieval requirements were changed. On the other hand, for participants in the Related Lure condition, the initial conceptual-based criterion would be very liberal and the level of performance attainable by persisting in that criterion would be very poor (due to high levels of false alarms). As such, these participants learned to avoid this conceptual-based criterion and supposedly relied on a more distinctive or feature-based encoding and response strategy. When, in Experiment 1, the nature of the recognition test was changed for participants in the Unrelated Lure condition, such that all lures were from the same category as the study list, performance deteriorated and the response criterion became as liberal as the initial response bias exhibited by participants in the Related Lure condition. Although this difference in criterion from the first three to the last study–test cycle (in Experiment 1) was quite dramatic, we account for it in terms of the same conceptual-based encoding and response strategy that gave rise to very different consequences when the test requirements were changed. Moreover, we found more direct evidence for the hypothesised learning process of avoiding a conceptual-based strategy, which occurred for participants in the Related Lure condition, as suggested by the linear of gradually reducing bias for high centrality items. It is also noteworthy that this effect emerged only for participants in the Related Lure condition (in both Experiments 1 and 2). Also, we were able to show that the avoidance of a knowledge or conceptual-based strategy and the development of an item or distinctive-based strategy gave an advantage on the free recall test to participants in the Related Lure condition, compared to participants in the Unrelated Lure condition, who supposedly did not learn to avoid the thematic or conceptual-based strategy (Hunt & McDaniel, 1993; Nairne, 2006). Finally, correlations between d′ and c scores from the three first study–test cycles and the number of items recalled at the last study–test cycle (free recall test) provided evidence that converged with the idea that the nature of the lures included in the successive recognition tests induced different encoding and processing strategies that benefited or hindered performance on the free recall test. The same category included in the recognition tests given to participants in Related Lure condition induced them to adopt strategies based on item distinctive features. These feature-based strategies apparently enhanced performance at free recall because the d′ scores and the number of items recalled were always strongly associated for these participants. On the contrary, only the c score from the third study–test cycle predicted the level of free recall for participants in the Unrelated Lure condition, showing a negative association. Thus, the more stringent was the criterion adopted by participants in the Unrelated Lure condition at the recognition test that preceded the last study–test cycle, the worse their performance was on the free recall test. As we explained before, we took stringency of criterion as an indication of knowledge or relational encoding and response strategy. Thus this inverse relationship corroborates the idea that the adoption of a relational strategy hinders free recall of related lists (Hunt & McDaniel, 1993; Nairne, 2006). Taken together these results point to a mediated test benefit—the ability to adapt encoding strategies to the nature of previous materials and retrieval tasks. This retrieval-induced strategy adaptation seems to be a valuable process to potentiate learning of related but different sets of items. Since we explored how testing potentiated the learning of new items and not how testing potentiated the learning of previously unrecalled but studied items (e.g., Karpicke, 2009), our proposed explanation of potentiation is not incompatible with other explanations such as the ones based on the activation of a specific search-set during retrieval that would lead to a better subsequent encoding (Grimaldi & Karpicke, 2012).

However, we think we must address three caveats before we conclude. The first is related to the suggestion made by Zaromb and Roediger (2010) that the power of repeated testing derives from the increased use of an organisational or conceptual-based strategy. In fact, Zaromb and Roediger showed that repeated testing led to better recall and to higher recall output organisation than repeated studying. Thus, if repeated testing leads to reliance on a more relational, conceptual-based or organisational strategy, how is it possible that our participants in the Related Lure condition became less reliant on it across study–test cycles? One possible answer is that the organisational benefits found by Zaromb and Roediger occurred from repeated testing of the same list, whereas in our case, a new study list was presented in each of the study–test cycles. This constant change of list may have made a specific conceptual-based strategy difficult to adopt and the poor diagnosticity of general conceptual-based strategies particularly easy to detect. Future research will be needed to clarify whether in study–test cycles with the same category conceptual-based strategies are avoided or used by participants that have studied the lists under conditions similar to our Related Lure condition.

The second caveat has to do with our preference for talking about an encoding and retrieval strategy. Why don’t we simply argue that our participants learned different response strategies during the performance of recognition tests with different test structures, while the encoding strategy remained invariant across conditions? We concur that this account can easily fit the results from Experiment 1. However, in Experiment 2, we changed the fourth test for both conditions, such that the specific requirement learned from previous recognition tests could no longer be used by any participant, independently of their condition, and we still obtained the predicted difference across conditions. Such a difference is typically obtained when the encoding set favours either relational conceptual-based or distinctive feature-based encoding (Hunt & McDaniel, 1993; Nairne, 2006). We agree that our paradigm cannot completely disentangle encoding and retrieval processes, but neither is it our aim (which is more how the interplay between these two processes can promote adaptation to a context with a given structure). Thus we contend that our proposal of an adaptive change of both encoding and retrieval strategies as a function of requirements of the relevant retrieval context fits better with the data from the two experiments.

The final and last caveat has to do with an alternative but related possible account, the *desirable difficulties in learning* framework from Bjork (1994a, 1994b, 2013). According to Bjork (2011), manipulations that make successful learning more difficult during study (introducing interference, spacing learning or using testing instead of restudy) often enhance long-term retention and transfer. In our case, although we always used similar study lists and standard recognition tests for both conditions, tests for the Unrelated Lure condition were patently easier, and thus the poorer performance of participants in the Unrelated Lure condition at free recall (Experiment 2) may simply indicate that the difficulties felt by participants in the Related Lure condition were indeed desirable. Although we do concur with this assessment, we believe that every encoding strategy has inevitable associated costs that become noticeable only under certain circumstances, that is, when the retrieval requirements change in appropriate ways. Note that whereas it is certainly true that participants who faced desirable difficulties (i.e., participants in the Related Lure condition) did perform better at the final study–test cycle in both experiments, it is very likely that participants in the Unrelated Lure condition would continue to outperform their colleagues if the final test had remained unchanged. Thus, we believe that although some difficulties in learning can be desirable in most long-term retention or transfer cases, it should always be possible to find a specific retrieval context in which the desirable difficulties cease to be desirable. Only future research can provide further insight into these questions.

We began by quoting John Anderson who argued that human memory did not evolve to give optimal performance in our silly laboratory tasks. And we agree. But we would add that even our silly laboratory tasks might lead our memories to "evolve" in adaptive ways to these tasks if we allow them to.

REFERENCES

Anderson, J. R. (1990). *The adaptive character of thought*. Hillsdale, NJ: Erlbaum.

Anisfeld, M., & Knapp, M. (1968). Association, synonymity and directionality in false recognition. *Journal of Experimental Psychology*, *77*, 171–179. doi:10.1037/h0025782

Arnold, K. M., & McDermott, K. B. (2013). Test-potentiated learning: Distinguishing between direct and indirect effects of tests. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*, 940–945. doi:10.1037/a0029199

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

Bjork, R. A. (1975). Retrieval as a memory modifier. In R. L. Solso (Ed.), *Information processing and cognition: The Loyola symposium* (pp. 123–144). Hillsdale: Erlbaum.

Bjork, R. A. (1994a). Memory and metamemory considerations in the training of human beings. In J. Metcalfe & A. Shimamura (Eds.), *Metacognition: Knowing about knowing* (pp. 185–205). Cambridge, MA: MIT Press.

Bjork, R. A. (1994b). Institutional impediments to effective training. In D. Druckman & R. A. Bjork (Eds.), *Learning, remembering, believing: Enhancing human performance* (pp. 295–306). Washington, DC: National Academy Press.

Bjork, R. A. (2011). On the symbiosis of learning, remembering, and forgetting. In A. S. Benjamin (Ed.), *Successful remembering and successful forgetting: A festschrift in honor of Robert A. Bjork* (pp. 1–22). London, UK: Psychology Press.

Bjork, R. A. (2013). Desirable difficulties perspective on learning. In H. Pashler (Ed.), *Encyclopedia of the mind*. Thousand Oaks, CA: Sage Reference.

Bjork, R. A., & Bjork, E. L. (1992). A new theory of disuse and an old theory of stimulus fluctuation. In A. Healy, S. Kosslyn, & R.' Shiffrin (Eds.), *From learning processes to cognitive processes: Essays in honor of William K. Estes* (Vol. 2, pp. 35–67). Hillsdale, NJ: Erlbaum.

Brainerd, C. J., & Reyna, V. F. (1988). Memory loci of suggestibility development: Comment on Ceci, Ross, and Toglia. *Journal of Experimental Psychology: General*, *117*, 197–200. doi:10.1037/0096-3445.117.2.197

Brainerd, C. J., & Reyna, V. F. (2001). Fuzzy-trace theory: Dual processes in memory, reasoning, and cognitive neuroscience. In H. W. Reese & R. Kail (Eds.), *Advances in child development and behavior* (Vol. 6, pp. 41–100). San Diego, CA: Academic Press.

Bruner, J. S., Goodnow, J. J., & Austin, G. A. (1956). *A study of thinking*. New York, NY: Wiley.

Connor, J. M. (1977). Effects of organization and expectancy on recall and recognition. *Memory & Cognition*, *5*, 315–318. doi:10.3758/BF03197576

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

d'Ydewalle, G. (1981). Test-expectancy effects in free recall and recognition. *The Journal of General Psychology*, *105*(2), 173–195. doi:10.1080/00221309.1981.9921071

Ebbinghaus, H. (1885). *Über das Gedächtnis* [Memory: A contribution to experimental psychology]. Leipzig: Duncker & Humblot.

Einstein, G. O., & Hunt, R. R. (1980). Levels of processing and organization: Additive effects of individual-item and relational processing. *Journal of Experimental Psychology: Human Learning & Memory*, *6*, 588–598. doi:10.1037/0278-7393.6.5.588

Finley, J. R., & Benjamin, A. S. (2012). Adaptive and qualitative changes in encoding strategy with experience: Evidence from the test-expectancy paradigm. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *38*, 632–652. doi:10.1037/a0026215

Glanzer, M., & Adams, J. K. (1985). The mirror effect in recognition memory. *Memory & Cognition*, *13*, 8–20. doi:10.3758/BF03198438

Greenberg, M. S., & Bjorklund, D. F. (1981). Category centrality in free recall: Effects of feature overlap or differential category encoding? *Journal of*

Experimental Psychology: Human Learning & Memory, *7*, 145–147. doi:10.1037/0278-7393.7.2.145
Grimaldi, P. J., & Karpicke, J. D. (2012). When and why do retrieval attempts enhance subsequent encoding? *Memory & Cognition*, *40*, 505–513. doi:10.3758/s13421-011-0174-0
Hall, J. W., Grossman, L. R., & Elwood, K. D. (1976). Differences in encoding for free recall vs. recognition. *Memory & cognition*, *4*, 507–513. doi:10.3758/BF03213211
Higham, P. A., & Brooks, L. R. (1997). Learning the experimenter's design: Tacit sensitivity to the structure of memory lists. *The Quarterly Journal of Experimental Psychology Section A*, *50*, 199–215. doi:10.1080/027249897392297
Hintzman, D. L. (1986). "Schema abstraction" in a multiple-trace memory model. *Psychological Review*, *93*, 411–428. doi:10.1037/0033-295X.93.4.411
Hunt, R. R., & McDaniel, M. A. (1993). The enigma of organization and distinctiveness. *Journal of Memory & Language*, *32*, 421–445. doi:10.1006/jmla.1993.1023
Izawa, C. (1971). The test trial potentiating model. *Journal of Mathematical Psychology*, *8*, 200–224. doi:10.1016/0022-2496(71)90012-5
Kantner, J., & Lindsay, D. S. (2010). Can corrective feedback improve recognition memory? *Memory & Cognition*, *38*, 389–406. doi:10.3758/MC.38.4.389
Karpicke, J. D. (2009). Metacognitive control and strategy selection: Deciding to practice retrieval during learning. *Journal of Experimental Psychology: General*, *138*, 469–486. doi:10.1037/a0017341
Karpicke, J. D., & Roediger, H. L., III. (2008). The critical importance of retrieval for learning. *Science*, *319*, 966–968. doi:10.1126/science.1152408
Landauer, T. K., & Bjork, R. A. (1978). Optimum rehearsal patterns and name learning. In M. M. Gruneberg, P. E. Morris, & R. N. Sykes (Eds.), *Practical aspects of memory* (pp. 625–632). London: Academic Press.
Leonard, J. M., & Whitten, W. B., II. (1983). Information stored when expecting recall or recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *9*, 440–455. doi:10.1037/0278-7393.9.3.440
Loftus, E. F. (1975). Leading questions and the eyewitness report. *Cognitive Psychology*, *7*, 560–572. doi:10.1016/0010-0285(75)90023-7
Lundeberg, M. A., & Fox, P. W. (1991). Do laboratory findings on test-expectancy generalize to classroom outcomes? *Review of Educational Research*, *61*, 94–106. doi:10.3102/00346543061001094
Macmillan, N. A., & Creelman, C. D. (2005). *Detection theory: A user's guide*. Mahwah, NJ: Erlbaum.
Maisto, S. A., DeWaard, R. J., & Miller, M. E. (1977). Encoding processes for recall and recognition: The effect of instructions and auxiliary task performance. *Bulletin of the Psychonomic Society*, *9*(2), 127–130. doi:10.3758/BF03336951
McClelland, J. L., & Chappell, M. (1998). Familiarity breeds differentiation: A subjective-likelihood approach to the effects of experience in recognition memory. *Psychological Review*, *105*, 724–760. doi:10.1037/0033-295X.105.4.734-760
Mccloskey, M. E., & Glucksberg, S. A. M. (1978). Natural categories: Well defined or fuzzy sets? *Memory & Cognition*, *6*, 462–472. doi:10.3758/BF03197480
McDaniel, M. A., Blischak, D. M., & Challis, B. (1994). The effects of test expectancy on processing and memory of prose. *Contemporary Educational Psychology*, *19*, 230–248. doi:10.1006/ceps.1994.1019
Meyer, G. (1934). An experimental study of the old and new types of examination: The effect of the examination set on memory. *Journal of Educational Psychology*, *25*, 641–661. doi:10.1037/h0073102
Meyer, G. (1936). The effect of recall and recognition on the examination set in classroom situations. *Journal of Educational Psychology*, *27*, 81–99. doi:10.1037/h0059333
Murphy, G. L. (2002). *The big book of concepts*. Cambridge, MA: MIT Press.
Nairne, J. S. (2006). Modeling distinctiveness: Implications for general memory theory. In R. R. Hunt & J. Worthen (Eds.), *Distinctiveness and memory*. New York, NY: Oxford University Press.
Neely, J. H., & Balota, D. A. (1981). Test-expectancy and semantic-organization effects in recall and recognition. *Memory & Cognition*, *9*, 283–300. doi:10.3758/BF03196962
Neisser, U. (1978). Memory: What are the important questions? In M. M. Grunenberg, P. E. Morris, & R. E. Sykes (Eds.), *Practical aspects of memory* (pp. 3–24). London: Academic Press.
Pinto, A. C. (1992). *Categorização de itens verbais: Medidas de frequência de produção e de tipicidade* [Categorization of verbal items: Measures of production frequency and typicality]. Porto: Relato Técnico de Centro de Psicologia Cognitiva.
Posner, M. I., & Keele, S. W. (1968). On the genesis of abstract ideas. *Journal of Experimental Psychology*, *77*, 353–363. doi:10.1037/h0025953
Roediger, H. L., III, & Butler, A. C. (2011). The critical role of retrieval practice in long-term retention. *Trends in Cognitive Sciences*, *15*, 20–27. doi:10.1016/j.tics.2010.09.003
Roediger, H. L., III, & Karpicke, J. D. (2006). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science*, *17*, 249–255. doi:10.1111/j.1467-9280.2006.01693.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. doi:10.1037/0278-7393.21.4.803
Roediger, H. L., III, Putnam, A. L., & Smith, M. A. (2011). Ten benefits of testing and their applications to educational practice. *Psychology of Learning and Motivation*, *55*, 1–36.
Schmidt, S. R. (1983). The effects of recall and recognition test expectancies on the retention of prose. *Memory & Cognition*, *11*, 172–180. doi:10.3758/BF03213472
Seger, C. A. (1994). Implicit learning. *Psychological Bulletin*, *115*, 163–196. doi:10.1037/0033-2909.115.2.163
Shiffrin, R. M., & Steyvers, M. (1997). A model for recognition memory: REM - Retrieving effectively

from memory. *Psychonomic Bulletin & Review*, *4*, 145–166. doi:10.3758/BF03209391

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. doi:10.3758/BF03193473

Thiede, K. W. (1996). The relative importance of anticipated test format and anticipated test difficulty on performance. *The Quarterly Journal of Experimental Psychology Section A*, *49*, 901–918. doi:10.1080/713755673

Tversky, B. (1973). Encoding processes in recognition and recall. *Cognitive Psychology*, *5*, 275–287. doi:10.1016/0010-0285(73)90037-6

Underwood, B. J., & Freund, J. S. (1968). Errors in recognition learning and retention. *Journal of Experimental Psychology*, *78*, 55–63. doi:10.1037/h0026163

Whitten, W. B., & Leonard, J. M. (1980). Learning from tests: Facilitation of delayed recall by initial recognition alternatives. *Journal of Experimental Psychology: Human Learning and Memory*, *6*, 127–134. doi:10.1037/0278-7393.6.2.127

Wissman, K. T., Rawson, K. A., & Pyc, M. A. (2011). The interim test effect: Testing prior material can facilitate the learning of new material. *Psychonomic Bulletin & Review*, *18*, 1140–1147. doi:10.3758/s13423-011-0140-7

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