Dissociating early- and late-selection processes in recall: The mixed blessing of categorized study lists

Mehmet A. Guzel · Philip A. Higham

Published online: 16 January 2013
© Psychonomic Society, Inc. 2013

Abstract Two experiments are reported in which we used type-2 signal detection theory to separate the effects of semantic categorization on early- and late-selection processes in free and cued recall. In Experiment 1, participants studied cue–target pairs for which the targets belonged to two, six, or 24 semantic categories, and later the participants were required to recall the targets either with (cued recall) or without (free recall) the studied cues. A confidence rating and a report decision were also required, so that we could compute both forced-report quantity and metacognitive resolution (type-2 discrimination), which served as our estimates of early- and late-selection processes, respectively. Consistent with prior research, having fewer categories enhanced the early-selection process (in performance, two > six > 24 categories). However, in contrast, the late-selection process was impaired (24 > six = two categories). In Experiment 2, encoding of paired associates, for which the targets belonged to either two or 20 semantic categories, was manipulated by having participants either form interactive images or engage in rote repetition. Having fewer categories again was associated with enhanced early selection (two > 20 categories); this effect was greater for rote repetition than for interactive imagery, and greater for free recall than for cued recall. However, late selection again showed the opposite pattern (20 > two categories), even with interactive-imagery encoding, which formed distinctive, individuated memory traces. The results are discussed in terms of early- and late-selection processes in retrieval, as well as overt versus covert recognition.

Keywords Recall · Recognition · Metamemory · Categorization · Type 2 signal detection theory

It has been well documented that semantic or categorical relatedness between study items (e.g., study lists composed of exemplars from semantic categories) has a facilitative effect on recall, as compared to no relationship between items (Cofer, Bruce, & Reicher, 1966; Guerin & Miller, 2008; Howard & Kahana, 2002; Puff, Murphy, & Ferrara, 1977; Runquist, 1970; Tehan, 2010; Thompson, Hamlin, & Roenker, 1972; Tulving & Pearlstone, 1966; but see Puff, 1970). Tulving and Pearlstone, for instance, tested the memory performance of their participants after the participants had studied categorized lists. They manipulated the number of categories (one, two, and four), as well as the length of the lists (12, 24, and 48 words). Participants were either presented with the category names as a retrieval cue at the time of testing or they were not (cued- vs. free-recall groups).¹ Tulving and Pearlstone found that participants reported more targets when category names were present than when these cues were absent. Furthermore, the effect of cueing was even greater when the lists were longer. These findings, and others like them, are generally understood to result from categorization acting as a powerful cue and providing access to information in memory that would not otherwise be accessible.

In contrast, the effects of categorization on recognition have been more equivocal. Some early studies demonstrated either no effect on recognition (e.g., Bruce & Fagan, 1970; Kintsch, 1968) or a facilitative effect (e.g., Connor, 1977; D’Agostino, 1969; Mandler, 1972; Neely & Balota, 1981). More recently, however, several reports have indicated that categorization has a detrimental effect on recognition (e.g.,

M. A. Guzel · P. A. Higham (✉)
School of Psychology, University of Southampton, Highfield,
Southampton SO17 1BJ, UK
e-mail: higham@soton.ac.uk

¹ Tulving and Pearlstone’s (1966) methodology differed somewhat from many other cued-recall experiments in that their cueing was not item specific; that is, participants were provided with a single category name to cue several items rather than a single one. In contrast, many other cued-recall experiments have used item-specific cueing, in which each to-be-remembered item was primed with its own cue (or several cues).

Dewhurst, 2001; Hintzman, 1988; Koutstaal & Schacter, 1997; Shiffrin, Huber, & Marinelli, 1995). In many of the latter studies, the number of exemplars per category in the study list was manipulated. The general finding was that as the number of exemplars increased, there might be a slight increase in the hit rate (HR), but it was greatly overshadowed by the increase in the false alarm rate (FAR), resulting in reduced old/new discrimination.

Early- and late-selection processes

Memory can be described as involving both early and late quality-control processes (e.g., Halamish, Goldsmith, & Jacoby, 2012; Koriat, Goldsmith, & Halamish, 2008), and this distinction may be useful for understanding the dissociative effects that categorization has on recall versus recognition. For example, a retrieval cue such as a category name may be used in recall to facilitate access to information in memory—that is, by enhancing the early-selection process. Conversely, categorization's deleterious effect on recognition likely results from lures sharing features with multiple exemplars of the same category in memory, making late selection (discrimination) difficult (Arndt & Hirshman, 1998; Dewhurst, 2001).

Although recall clearly involves an early-selection process, a late-selection process is surely implicated as well. This late-selection process of recall is likely to be impaired by categorization, just as overt recognition is. Thus, according to this analysis, categorization may have opposing effects on overall recall performance: The first effect is enhancement of early selection, which is offset by the second effect, a detriment to late selection. Evidently, if such opposing influences exist in recall, the former early-selection benefit overshadows the detriment to late-selection effectiveness, evinced by the fact that categorization typically improves recall. However, that does not mean that late-selection deficits are absent. The problem is that all that researchers have typically observed has been overall recall performance. What is needed is a methodology to separate the late-selection effects on recall from the early ones. In the next section, we describe such a methodology.

Separating early- and late-selection processes in recall

A great deal of metamemory research has investigated how people strategically regulate memory accuracy. Two frameworks have been utilized for this purpose. The first is Koriat and Goldsmith's (1996) framework that incorporates the *quantity-accuracy profile* (QAP) methodology (e.g., Halamish et al., 2012; see Goldsmith & Koriat, 2008, for a review). The second is *type-2 signal detection theory* (SDT),

advocated by Higham and colleagues (e.g., Higham, 2002, 2007; Higham & Arnold, 2007; Lueddeke & Higham, 2011; see also Higham, 2011, and Goldsmith, 2011, for discussions of the various merits and drawbacks of each approach). Both frameworks were originally developed to evaluate performance on tasks that incorporated a *report option* (the option to pass or withhold responses) that might be used to regulate memory accuracy. Both assumed that in response to some input (e.g., a question stem or a retrieval cue), a search of memory ensues. This search potentially yields a number of candidate answers that are subject to a late-selection process (i.e., they are monitored for accuracy). If a candidate reaches some criterion of acceptability on the basis of output from a monitoring mechanism that assigns confidence in the correctness of the candidate answer, the candidate is reported. Otherwise, it is withheld.

By strategically using the report option to filter out unwanted candidate responses, the accuracy of the information that is reported is likely to be enhanced relative to cases in which no such option exists. However, for this enhancement to occur, *resolution* (the ability to discriminate between one's own correct and incorrect answers) must be above chance. In the type-2 SDT framework, the framework adopted in the present research, resolution is measured by a discrimination index such as d' . Whichever discrimination measure is used, it is based on the type-2 HR and FAR, which are the proportions of correct versus incorrect answers, respectively, that are reported. To determine these rates, it is necessary to know the answers that were reported, but also those that were withheld. Of course, in most scenarios, withheld answers are the participants' private knowledge. However, in studies on strategic accuracy regulation, it is possible to obtain these answers by forcing output in a separate phase of the experiment (e.g., Koriat & Goldsmith, 1996) or by asking participants to answer all questions but to indicate whether they would like to report or withhold each answer (Higham, 2007, Exp. 2). By doing so, a measure of participants' ability to discriminate their own correct answers from their incorrect ones can be derived. This measure—resolution or type-2 discrimination—is what particularly can be used to estimate the late-selection process in recall.

In addition to resolution, several other performance measures can be derived using the type-2 SDT framework. The first of these is *forced-report quantity*, the proportion of correct answers on the test after all questions have been answered, ignoring any report/withhold decisions, designated by the parameter f (Higham, 2007). In the context of recall, forced-report quantity is likely to be the most sensitive to the early-selection process, because it is not affected by any late-selection withholding of unwanted candidates. Consequently, any facilitative effects derived from variables such as categorization should be fully realized in this measure.

A third performance measure is *bias*. Just as with standard SDT, bias reflects the overall tendency to respond “yes” or “no,” independent of discrimination ability. In the type-2 context that incorporates a report option, bias reflects the extent to which a participant is willing to report answers. At forced report, report bias is maximally liberal, whereas if participants withheld all candidate responses, report bias would be maximally conservative. Bias can be measured in a number of different ways with indices such as C or B''_D (Donaldson, 1992).

Finally, the proportion of correct answers after unwanted (withheld) ones are filtered out (henceforth, *free-report quantity*) is a fourth performance measure. This measure is the one that has typically been used in many recall studies. That is, participants are usually instructed to provide as many targets as possible (to explicit cues in cued recall or to personal cues in free recall) but to avoid guessing, which effectively provides them with a report option. Although this measure is typically the one adopted in recall studies, it is probably the most complex and heterogeneous of all of the measures so far described, because, as we will show, it is a hybrid measure that is sensitive to resolution (late selection), forced-report quantity (early selection), and report bias.

Using type-2 signal detection analyses, Higham and Tam (2005) were able to separately estimate the early- and late-selection stages of cued recall. In their second study, participants studied a mixed list of strongly and weakly associated cue–target pairs (e.g., *homicide*–*MURDER* and *bats*–*BLOOD*, respectively). Participants then took a cued-recall test and were given the opportunity to earn points if they were confident in their response to the cue, or to advance to a guessing stage if they were not. If a correct cued-recall response was offered for points, it was considered as being analogous to reporting, and participants’ cumulative point total increased, whereas the point total decreased if the response was incorrect. Responses offered as guesses were deemed withheld and had no effect on the cumulative point total, regardless of their accuracy.

The cues from Higham and Tam’s (2005) second experiment that were of particular relevance to the present discussion were the studied cues from weakly associated and strongly associated study pairs. Whereas forced-report quantity was higher for strong studied cues than for weak studied cues, the opposite was true for resolution. That is, Higham and Tam’s type-2 signal detection measure of resolution, A' , was .87 for weak cues, but only to .74 for strong cues. Higham and Tam reasoned that, whereas strong cues were likely to facilitate the early-selection process and lead to accessing the target in memory, the relatedness between the target and the other retrieved candidates in the postretrieval

search set was higher for strong than for weak cues, interfering with late selection and resulting in poorer resolution performance. For example, in response to the strong cue *homicide*, the candidates *MURDER* (target), *death*, *kill*, and *die* may have been considered for report. These candidates are not just related to the cue (*homicide*), but are related to each other as well. Thus, monitoring the retrieved candidates would have been difficult, despite the fact that the target was likely to have been amongst them. In contrast, consider the search set for a weak cue such as *bats*. A weak cue such as this might produce candidates such as *vampire*, *blind*, *cave*, and *fangs*, because of associative relationships, as well as the target *BLOOD*, via conscious recollection. Because *BLOOD* is only weakly related to both the cue and the other generated candidates, monitoring the target when it was in the search set was made easier. Hence, candidate interrelatedness and categorization assisted in the early-selection retrieval process, which produced a high forced-report quantity (f), but impaired the late-selection monitoring process, which produced low resolution (A').

Present research

Higham and Tam’s (2005) research supports the idea that categorization has opposing effects on early- and late-selection processes of recall. However, those researchers did not directly manipulate the number of categories in the study list, and their account of the resolution/quantity dissociation was speculative. To address the issue more directly, in the present research we experimentally manipulated the categorical structure of the study lists and examined the effect on recall performance in two experiments that included a report/pass option and confidence ratings. The approach that we adopted is relatively novel, and little is known about how early- and late-selection processes might compare if experimenter-defined cues were presented at test (cued recall) versus if participants were required to perform a self-initiated search of memory using their own cues (free recall). Consequently, we tested both free- and cued-recall groups in both experiments. On the basis of the reasoning above, we expected to find that categorization would produce dissociative effects on early-selection (forced-report quantity) versus late-selection (resolution) processes. Furthermore, by using a type-2 methodology, we demonstrated that free-report quantity, the typical performance measure used in many recall studies, is a heterogeneous index of memory, being influenced by both early- and late-selection processes, as well as by report bias. In line with this finding, we

demonstrated that the effect of categorization on free-report quantity is not as great as the effect on forced-report quantity, presumably because the late-selection process impaired the former measure but not the latter. Together, the results led us to conclude that previous demonstrations of the facilitative effect of categorization on the early-selection process of recall have underestimated the true effect of categorization on memory access.

In Experiment 1, we manipulated the structure of the study list, such that it contained exemplars from two, six, or 24 different semantic categories. We predicted that having fewer categories would increase early-selection forced-report quantity, because the lists would be more cohesive, but would impair late-selection resolution, because of increased confusability amongst the candidate answers. In Experiment 2, we again explored the effect of categorization on the early- and late-selection processes of recall, but also investigated the moderating effect that trace individuation and distinctiveness (caused by interactive-imagery instructions) might have on the categorization effects.

Experiment 1

The participants in Experiment 1 studied three study lists consisting of 24 cue–target word pairs, and each list was followed by either a cued- or a free-recall test. There were no associations between the cues and targets or between the cues, but critically, the number of categories to which the targets belonged across the list as a whole was manipulated (two, six, or 24).

We hypothesized that forced-report quantity would be inversely related to the number of categories in the study list. That is, it would show the pattern $two > six > 24$ categories, because of an early-selection advantage with categorized lists. However, because categorization is likely to produce a post-retrieval search set of highly confusable candidates, and because this confusability is likely to worsen as the number of exemplars belonging to the category increases, we predicted the opposite pattern for resolution performance. That is, resolution would show the pattern $24 > six > two$, producing a retrieval/monitoring dissociation.

Method

Participants A group of 60 students at the University of Southampton participated in the study for course credit or £5 payment. The participants, who spoke English as a first language, were randomly assigned to one of two groups: cued recall ($n = 30$; age: $M = 19.6$ years, $SEM = 0.69$) or free recall ($n = 30$; age: $M = 19.2$ years, $SEM = 0.27$).

Design and materials A 2 (recall type: cued vs. free) $\times 3$ (category number: two vs. six vs. 24) mixed factorial design was used, with Recall Type as the between-subjects factor.

Three study lists consisting of 24 cue–target word pairs were created. All of the lists had unrelated cue–target pairs (e.g., victim–TEACHER, impact–MANAGER). However, the lists consisted of target words that were exemplars selected from two, six, or 24 discrete semantic categories. To achieve this list structure, a target pool of 24 mutually exclusive categories consisting of 13 exemplars each (312 targets in total) were selected from Van Overschelde, Rawson, and Dunlosky (2004). A total of 72 additional words were selected to form a pool of items to act as cues. According to the University of South Florida Free Association Norms (Nelson, McEvoy, & Schreiber, 1998), the cue words had negligible semantic (as well as categorical) association with each other or with any of the target words.

The study lists were created individually by software for each participant prior to the experiment. The program first selected 12 exemplars from two of the 24 available categories (targets pool). Both the categories chosen and the exemplar selections within categories were random. These words constituted the targets for the two-category list. Then, the program randomly selected four exemplars from six of the remaining 22 categories. These words were used as target words for the six-category list. Finally, the program randomly selected one exemplar from each of the 24 categories to create the 24-category (unrelated) list. For the categories that had been used to create the two-category list, the 13th exemplar was used for the 24-category list. For the categories used to create the six-category lists, one exemplar was chosen from amongst the nine remaining exemplars. For all of the other categories, one word was randomly selected from the 13 available exemplars. This procedure ensured that none of the lists (two-category, six-category, or 24-category) consisted of the same categories or target words. Finally, the 72 words in the cues pool were randomly allocated to each of the 72 targets across the three lists, to make 72 cue–target pairs (3×24 pairs). The cue assignments were freshly randomized for each participant.

Procedure Participants were tested individually in a quiet and dimly lit laboratory. They were tested in three computerized study–test cycles. In each study phase, the participants were presented with the two-category, six-category, or 24-category study list. The presentation order of the lists was counterbalanced, which resulted in six versions (two–six–24, six–24–two, 24–two–six, two–24–six, six–two–24, and 24–six–two). The participants were randomly assigned to one of the versions, and the presentation order of pairs within each list was freshly randomized for each participant.

The participants started with a practice study–test cycle with a study list of five word pairs (which were different from the other items used in the experiment). After the practice phase, the participants were warned that they were about to start the actual study in which their responses would be scored. In each study phase, the participants were presented with a study list of 24 word pairs in a random order, and they were instructed that they would be responsible for remembering and reporting the target words—that is, the words on the right-hand side in the pairs. Participants started the presentation of the lists by clicking on a “Start the presentation” button located on the computer screen. Each word pair was presented on the computer screen for 3 s, with a 1 s interstimulus interval (ISI). The cue–target pairs were presented in capital letters and separated by a hyphen between the words (e.g., “EFFORT - UNCLE”). Following the presentation of each list, each participant solved some algebra calculations of moderate difficulty and/or Sudoku puzzles for 5 min as a distractor activity.

All testing was completed on a computer. Cued-recall participants were asked to type in the targets in a Targets column next to cues provided in a Cues column. All of the cues were those previously studied. Participants could provide targets to the cues in any order that they wished, but omissions were not permitted. A checkbox was displayed next to each target, which was used to indicate whether or not the response was to be reported. In particular, participants checked the Report checkbox to indicate that they felt confident enough to report that answer, whereas they checked the Pass checkbox if they did not. The report-option checkbox could be used to distinguish between responses that would likely have been provided on a typical cued-recall test and responses that would have been withheld. Finally, participants rated each response in terms of how confident they were that the response was correct, on a seven-point scale provided next to each response (1 = *Not at all confident correct*, 4 = *Fairly confident correct*, 7 = *Completely confident correct*). The presentation order of the cues was freshly randomized for each participant. At the end of the testing phase, the participants were given a written debriefing form and the researcher responded to any queries. The study lasted 50–60 min.

The testing procedure for the free-recall participants was exactly the same as for the cued-recall group, except that these participants were not given any cue words. Instead, they were asked to write down the target words in 24 empty spaces under a Targets column displayed on the computer screen. As with the cued-recall participants, a response was required in every space, even if they had to guess, but again the report-option checkboxes were available to distinguish between items that would likely have been provided on a typical free-recall test and items that would have been withheld.

Results

Separate 2×3 mixed analyses of variance (ANOVAs) are reported below, each testing for effects of categorization (two vs. six vs. 24 categories; within subjects) and recall type (cued vs. free recall; between subjects) on three different performance measures: (1) forced-report quantity, our measure of the early-selection process; (2) resolution (indexed by the confidence ratings), our measure of the late-selection process; and (3) the difference between free- and forced-report quantity, which should reflect the interplay of early- and late-selection processes.

Forced-report quantity Early-selection recall performance was indexed by forced-report memory quantity: that is, the proportion of targets out of the 24 presented in each study list that were recalled, regardless of the report/pass decision. For the cued-recall group, scoring was liberal (i.e., it did not matter whether or not a target response was paired with the correct cue). Strict scoring produced results very similar to those from liberal scoring.

The 2×3 mixed ANOVA conducted on forced-report quantity (see Fig. 1) revealed a main effect of the category number, $F(2, 116) = 58.72$, $p < .001$, $\eta^2 = .503$. Follow-up analyses revealed that quantity was significantly higher for the two-category condition ($M = .62$, $SEM = .02$) than for the six-category condition ($M = .43$, $SEM = .03$), $t(59) = 7.11$, $p < .001$, which in turn was significantly higher than for the 24-category condition ($M = .35$, $SEM = .03$), $t(59) = 3.12$, $p = .003$. Quantity in the two- and 24-category conditions was also significantly different, $t(59) = 9.49$, $p < .001$. However, this main effect was qualified by an interaction with recall type, $F(2, 116) = 6.64$, $p = .002$, $\eta^2 = .10$. The interaction occurred because the increase in retrieval performance as the number of categories reduced was greater in the free- than in the cued-recall group (see Fig. 1), although both effects were significant, $F(2, 58) = 41.59$, $p < .001$, $\eta^2 = .59$, and $F(2, 58) = 17.64$, $p < .001$, $\eta^2 = .38$, respectively. The main effect of recall type was not significant, $F < 1$.

Resolution A common measure of resolution is the Goodman–Kruskal gamma correlation coefficient. However, gamma has been shown to have a number of problems (see Masson & Rotello, 2009; Rotello, Masson, & Verde, 2008), such as being influenced by response bias. As a preferable alternative, the receiver operating characteristic curves used in signal detection analysis can be used to derive a measure of monitoring (see, e.g., Luna, Higham, & Martin-Luengo, 2011). In particular, the area under the (receiver operating characteristic) curve (AUC) has been described by Macmillan and Creelman (2005) as a “good index of sensitivity” (p. 64), and it can easily be computed using the trapezoidal rule (see, e.g., Green & Moses, 1966; Pollack, Norman,

Fig 1 Mean free- and forced-report quantities in the cued- and free-recall groups of Experiment 1, as a function of number of categories in the study lists (two vs. six vs. 24). Error bars represent standard errors of the means

& Galanter, 1964). Chance discrimination and perfect discrimination correspond to AUCs of .5 and 1.0, respectively.

AUC was computed for each participant and subjected to a 2×3 mixed ANOVA (see Table 1).² The main effect of the number of categories in the study lists was significant, $F(2, 108) = 4.07, p = .02, \eta^2 = .07$: Participants monitored their responses *worse* as the list contained fewer categories. Specifically, resolution was better for the 24-category list ($M = .94, SEM = .01$) than for the two-category list ($M = .89, SEM = .01, t(55) = 3.07, p = .003$, and also better than for the six-category list ($M = .89, SEM = .02, t(55) = 2.34, p = .02$). The resolutions were comparable between the two-category and six-category lists, $t(59) = 0.08, p = .94$. Neither the main effect of recall type nor the interaction from the ANOVA was significant, largest $F(1, 54) = 2.70, p = .11$.

Free- versus forced-report quantity Free-report quantity, shown in Fig. 1, is the typical measure of cued-recall performance, in that participants are often requested to avoid guessing, which provides them with a report option. We

Table 1 Means (with standard deviations) for resolution (indexed by AUC) in the cued- and free-recall groups of Experiment 1, as a function of the number of categories in the study list (two, six, or 24)

Group	Number of Categories		
	2	6	24
Cued recall	.90 (.10)	.92 (.10)	.95 (.08)
Free recall	.87 (.10)	.86 (.21)	.92 (.11)

AUC = area under the (receiver operating characteristic) curve.

initially analyzed this measure as we had the others, in a 2×3 mixed ANOVA. The pattern of significant and non-significant main effects and interactions was the same that had been seen in the analysis on forced-report quantity reported above, so the details will not be reported here, in the interest of brevity.

Of greater interest was any *advantage* that forcing output had on the quantity score. This difference, particularly, is what is theoretically likely to increase with the poorer resolution induced by categorization. Furthermore, demonstrating this quantity-difference/resolution relationship suggests that studies that have employed free-report quantity as the measure of recall performance will likely have underestimated categorization's effect on the early-selection recall process. Consistent with the resolution results, a 2×3 mixed ANOVA on the quantity difference scores yielded a main effect of category number, $F(2, 116) = 10.04, p < .001, \eta^2 = .15$. Pairwise comparisons indicated that the quantity difference in the two-category condition ($M = .07, SEM = .02$) was greater than the difference in the six-category condition ($M = .04, SEM = .01, t(59) = 2.19, p = .03$, and also greater than the difference in the 24-category condition ($M = .02, SEM = .01, t(59) = 4.73, p < .001$). The quantity differences in the two- and six-category conditions were also significantly different, $t(59) = 2.24, p = .03$. No other main effect or interaction from the ANOVA was significant, largest $F(2, 116) = 1.72, p = .18$.

Discussion

Experiment 1 confirmed the hypothesis that fewer categories in the study list would enhance the early-selection process of recall (forced-report quantity), while decreasing the late-selection process (resolution). These results are consistent with those of Higham and Tam (2005), who also found a dissociation between memory quantity and resolution in cued recall using strong and weak cues, and who suggested that the dissociation was caused by category effects in the search set. However, they did not directly manipulate the categorical structure of the study lists, and their interpretation was speculative. In contrast, categorical structure was directly manipulated in the present

² Four participants were dropped from this analysis, one because of failure to retrieve any targets in one of the experimental conditions (and, hence, no hit rate), and three additional ones because of extremely low, below-chance monitoring scores (near zero), suggesting that the confidence scale was being used in the opposite way from what had been instructed (i.e., high confidence was being represented by low rather than high values on the scale).

experiment, and the results provided solid evidence that categorization has opposing effects on early- versus late-selection processes in recall.

The offsetting effect of impaired late selection was brought to the fore in the analysis of free-report quantity, the measure of recall used in most studies. This measure was not as sensitive to categorization as was forced-report quantity; in other words, the difference between free- and forced-report quantities increased as the number of categories in the study list decreased (Fig. 1). We reasoned that this was due to the impurity of the free-report measure; although early-selection processes influence it, so do late-selection processes. Because these processes work in opposition to each other, the result is that free-report quantity has decreased sensitivity to categorization's effect.

Although the dissociative pattern was observed in both cued and free recall (Fig. 1), it is worth noting that the effect of category number on forced-report quantity was larger for free than for cued recall. We attribute this enhanced effect to the absence of any overt cues in free recall. Without cues being provided by the experimenter to aid retrieval, participants may have been more reliant on using target categorical structure to cue targets in memory, thus boosting any effects of categorization.

Experiment 2

Experiment 1 demonstrated that categorized study lists had a dissociative effect on forced-report quantity (early selection) versus resolution (late selection). In Experiment 2, we tested the generality of this dissociation by varying the distinctiveness of the encoding. We reasoned that if the experimental circumstances encouraged participants to form distinctive, individuated memory traces, the early-selection process would likely be facilitated, and the late-selection process might no longer suffer from intertarget similarity because the search sets would be of limited size. The upshot would be no dissociation between forced-report quantity and resolution, of the sort observed in Experiment 1.

To test this possibility, we had participants encode the cue-target pairs with *interactive imagery* and compared the performance with this type of encoding to performance using *rote repetition*. Previous research comparing these types of encoding had shown a substantial mnemonic benefit for interactive imagery as compared to rote repetition, presumably because imagery creates distinctive memory traces that are highly retrievable (e.g., Bower, 1970; Bower & Winzenz, 1970; Robbins, Bray, Irvin, & Wise, 1974). Consequently, we expected that, relative to rote-repetition encoding, interactive imagery would both facilitate the early-selection process and limit problems with late selection caused by categorization. The latter effect, in turn,

would result in a less pronounced dissociative pattern between forced-report quantity and resolution.

Method

Participants A total of 64 students at the University of Southampton, who spoke English as a first language, participated in the study. They were compensated for their time with either course credits or £5 payment, and were randomly allocated to one of the groups of the study: cued recall ($n = 32$; age: $M = 24.5$ years, $SEM = 0.99$) or free recall ($n = 32$; age: $M = 27.0$ years, $SEM = 0.97$).

Design and materials Because the largest differences in Experiment 1 were between the two- versus the 24-category (unrelated) conditions, we decided to eliminate the intermediate six-category condition from this experiment. Also, the study lists were shortened slightly, to 20 instead of 24 pairs. Hence, Experiment 2 had a 2 (recall type: cued vs. free) $\times 2$ (encoding strategy: interactive imagery vs. rote repetition) $\times 2$ (category number: two vs. 20) mixed factorial design, with recall type as the only between-subjects variable. The dependent variables were forced-report quantity (early-selection process), resolution (late-selection indexed by AUC), and free-report quantity.

Four study lists, composed of 20 cue–target word pairs each, were constructed. Two of the lists were two-category (T) target lists, in which the 20 targets belonged to two discrete, mutually exclusive categories of 10 exemplars each (T1 and T2) chosen from Van Overschelde et al. (2004). The T1 categories were fruits and animals, whereas the T2 categories were pieces of clothing and musical instruments. The targets within each two-category target list (T1 and T2) were all nouns and had comparable free-association means (i.e., percentages of participants giving the target in response to the category name). The remaining two lists of 20 cue–target pairs were multiple-category (M) target lists, in which the targets belonged to 20 different categories (i.e., the targets were unrelated). The 20 unrelated targets in M1 and M2 were composed of two subsets of 10 targets each that matched the two 10-target categories in T1 and T2 in terms of written frequency (Kučera & Francis, 1967), as well as in terms of imageability and concreteness values from the MRC Psycholinguistic Database. In other words, each subset of 10 targets in M1 was matched with each subset of 10 targets in T1 as closely as possible, and the same was true of the subsets within M2 and T2. This matching was done so that one version of the T and M lists (e.g., T1 and M1) could be studied with interactive imagery and the other version (e.g., T2 and M2) could be studied with rote repetition without concern for confounding stimulus attributes. Because it was difficult to match the subsets

on all dimensions simultaneously, written frequency and imageability values were given priority over concreteness, as these were considered more critical. Nonetheless, any failure to perfectly match the items across these stimulus attributes was not serious because of the counterbalancing procedure outlined below.

Cue words were chosen so that the targets from the T and M lists were weakly associated with them, but unlike the targets in the T lists (T1 and T2), they shared no relationship amongst themselves. Each subset of 10 cues within the T1, T2, M1, and M2 lists matched their corresponding targets in terms of written frequency, concreteness, and imageability values.

Procedure The participants were tested individually in four computerized study–test cycles. They studied two T lists and two M lists, with the different versions of each list being assigned a different encoding strategy, and the T and M lists alternating across the cycles. However, encoding strategy was blocked such that participants performed either interactive imagery for the first two study–test cycles, and rote repetition for the latter two, or vice versa. With these constraints, the order of the study lists was counterbalanced to produce 16 formats, with four participants assigned to each format.

Presentations of the study lists, distractor activities, and the procedure used in the testing phases, during which the responses were recorded electronically, were exactly the same as in Experiment 1. Unlike in Experiment 1, however, each word pair appeared on the computer screen for 4 s, with a 1 s ISI.³

Prior to the experiment, the participants had a practice phase in which they studied and recalled unrelated targets from two lists consisting of five unrelated cue–target pairs. The first list was encoded with interactive imagery, and the second with rote repetition. Whether or not the cue words were provided to the participant at the time of testing depended on the group to which the participant was allocated (cued vs. free recall). Once the participants finished the practice tests, they were warned that they were about to start the actual study, during which their responses would be counted. After the study was completed, the participants were given a written debriefing statement and the experimenter responded to their queries. As it was once again self-

paced during the testing phases, the study lasted between 55 and 65 min.

Results

Separate 2 (recall type: cued vs. free) $\times 2$ (category number: two vs. twenty) $\times 2$ (encoding strategy: interactive imagery vs. rote repetition) mixed ANOVAs analogous to those of Experiment 1, with Recall Type as the only between-subjects factor, were conducted on (1) forced-report quantity, (2) resolution (AUC), and (3) the difference between free- and forced-report quantities.

Forced-report quantity Figure 2 shows mean forced-report quantity, our measure of the early-selection process in recall, for the various conditions of Experiment 2. As in Experiment 1, scoring was liberal for the cued-recall group, which yielded results very similar to those from strict scoring.

The $2 \times 2 \times 2$ mixed ANOVA conducted on forced-report memory quantity yielded a main effect of the number of categories, $F(1, 62) = 126.08$, $p < .001$, $\eta^2 = .67$, and a main effect of encoding strategy, $F(1, 62) = 56.01$, $p < .001$, $\eta^2 = .48$. Forced-report quantity was significantly higher in the two-category condition ($M = .63$, $SEM = .02$) than in the 20-category condition ($M = .37$, $SEM = .02$), and it was higher with interactive imagery ($M = .59$, $SEM = .02$) than with rote repetition ($M = .41$, $SEM = .02$). However, both of these main effects were qualified by an interaction between the number of categories and encoding strategy, $F(1, 62) = 11.16$, $p = .001$, $\eta^2 = .15$. Pairwise comparisons showed that when the items were encoded with rote repetition, the difference between the two-category ($M = .56$, $SEM = .03$) and 20-category ($M = .26$, $SEM = .02$) conditions was greater than when interactive imagery was used ($M = .69$, $SEM = .03$, and $M = .49$, $SEM = .04$, respectively), although both effects were significant, $t(63) = 9.69$, $p < .001$, and $t(63) = 6.20$, $p < .001$, respectively.

Recall type interacted with the number of categories, $F(1, 62) = 34.73$, $p < .001$, $\eta^2 = .36$. Pair-wise mean comparisons indicated that the difference in recall between the two-category ($M = .68$, $SEM = .03$) and 20-category ($M = .29$, $SEM = .03$) conditions was greater in the free- than in the cued-recall group ($M = .58$, $SEM = .03$, and $M = .46$, $SEM = .03$, respectively), although both effects were significant, $t(31) = 12.86$, $p < .001$, and $t(31) = 3.57$, $p = .001$, respectively.

Recall type also interacted with encoding strategy, $F(1, 62) = 37.39$, $p < .001$, $\eta^2 = .38$. Pairwise comparisons showed that, whereas the cued-recall group had higher quantity when they used interactive imagery ($M = .68$, $SEM = .04$) than when they used rote repetition ($M = .35$,

³ We needed an optimal presentation rate (the longest one possible) that would allow participants to construct images in a given study time, but at the same time, would not have allowed them to shift from rote repetition to interactive imagery when they were repeating the word pairs out loud. It was suggested by John Dunlosky (J. Dunlosky, personal written communication, May 18, 2010) that any presentation time between 4 and 10 s is equally effective for constructing a visual image between pairs (see, e.g., Dunlosky & Hertzog, 1998). Therefore, as a slowest possible presentation rate to construct mental images, the word pairs were presented with a rate of 4 s with a 1 s ISI.

Fig 2 Mean free- and forced-report quantities in the cued-recall (panels A and B) and free-recall (panels C and D) groups in Experiment 2, as a function of number of categories in the study lists (two vs. 20) and encoding strategy (rote repetition [panels A and C] vs. interactive imagery [panels B and D]). Error bars represent standard errors of the means

$SEM = .03$), $t(31) = 7.46$, $p < .001$, the same difference was not significant in the free-recall group ($M = .50$, $SEM = .03$, and $M = .47$, $SEM = .02$, respectively), $t(31) = 1.669$, $p = .11$. No other main effect or interaction was significant from this analysis, largest $F(1, 62) = 1.13$, $p = .29$, $\eta^2 = .02$.

Resolution As in Experiment 1, we indexed resolution with AUC (see Table 2). It was not possible to compute AUC for eight participants (seven for cued recall and one for free recall) because of forced-report quantity scores of 1 or 0 in at least one experimental condition, rendering empty cell(s) in the analysis.

Table 2 Means (with standard deviations) for resolution (indexed by AUC) in the cued- and free-recall groups of Experiment 2, as a function of the number of categories in the study list (two or 20) and encoding type (rote repetition or interactive imagery)

Group	Number of Categories	
	2	20
Rote Repetition		
Cued recall	.64 (.28)	.88 (.18)
Free recall	.72 (.21)	.94 (.14)
Interactive Imagery		
Cued recall	.75 (.26)	.90 (.12)
Free recall	.69 (.25)	.89 (.13)

AUC = area under the (receiver operating characteristic) curve.

The $2 \times 2 \times 2$ mixed ANOVA conducted on the AUC scores revealed only a main effect of category number, $F(1, 54) = 131.82$, $p < .001$, $\eta^2 = .71$. As in Experiment 1, this main effect was in the opposite direction from that observed for quantity. That is, resolution was better in the 20-category condition ($M = .90$, $SEM = .02$) than in the two-category condition ($M = .70$, $SEM = .01$). No other main effect or interaction was significant from this analysis, largest $F(1, 54) = 2.40$, $p = .13$, $\eta^2 = .04$.

Free- versus forced-report quantity As in Experiment 1, we analyzed free-report quantity just as we had forced-report quantity, in a $2 \times 2 \times 2$ mixed ANOVA. The results of this ANOVA closely resembled those for forced-report quantity. That is, the same pattern of significant and nonsignificant main effects and interactions was obtained, so the full results of this analysis will not be reported, in the interest of brevity.

Of greater interest was the difference between free- and forced-report quantities, which would reflect the deleterious effect of an impaired late-selection process on free-report recall performance. Figure 2 suggests that, just as in Experiment 1, the quantity difference increased as the number of categories in the study list decreased. This observation was confirmed statistically with a $2 \times 2 \times 2$ mixed ANOVA on the quantity difference scores, which yielded a main effect of category number, $F(1, 62) = 26.05$, $p < .001$, $\eta^2 = .30$. The difference between free- and forced-report quantities was greater in the two-category condition ($M = .06$, $SEM = .01$) than in the 20-category condition ($M = .02$, $SEM = .00$). The

analysis also revealed a main effect of encoding type, $F(1, 62) = 5.44$, $p < .02$, $\eta^2 = .08$, which was qualified by an interaction between encoding type and recall type, $F(1, 62) = 5.44$, $p < .02$, $\eta^2 = .08$. Pairwise comparisons indicated that in the cued-recall group, the quantity difference score was significantly greater in the rote-repetition condition ($M = .06$, $SEM = .01$) than in the interactive-imagery condition ($M = .02$, $SEM = .01$), $t(31) = 2.52$, $p = .02$. However, this difference was not significant in the free-recall group (both rote repetition and interactive imagery: $M = .04$, $SEM = .01$), $t(31) = 0.00$, $p = 1.0$. No other main effect or interaction was significant from the ANOVA, largest $F(1, 62) = 2.20$, $p < .14$, $\eta^2 = .03$.

Discussion

As predicted, the early-selection process was facilitated by distinctive, interactive-imagery encoding. Forced-report quantity was higher with interactive-imagery encoding than with rote-repetition encoding, but this facilitative effect was only evident if the study cues were presented again at test (cued recall). This pattern demonstrates that the beneficial effects of interactive imagery are cue dependent; for example, interactive imagery does not cause targets to be encoded more deeply (Craik & Lockhart, 1972). Rather, it appears to enhance the efficacy of studied cues compared to rote repetition.

Categorization also had an effect on the early-selection process, but, unexpectedly, the type of encoding moderated its effect. In particular, categorization had a larger effect on forced-report quantity if the items were encoded with rote repetition than if they were encoded with interactive imagery. This pattern likely reflects the cue set that participants relied on in recall. That is, relative to rote-repetition encoding, interactive-imagery encoding may have biased participants to use the studied cues to prompt memory rather than the categorical structure, particularly in cued recall. In a similar vein, participants were more likely to show categorization effects on the early-selection process if they were engaged in free as opposed to cued recall, replicating an analogous effect in Experiment 1. Again, this pattern most likely emerged because, compared to free recall, the studied cues were the primary source of memory prompting in cued recall, rather than the categorical structure.

In terms of the late-selection process (resolution), we predicted that interactive imagery would create individuated, distinctive memory traces that would be less confusable than would the traces laid down by rote-repetition encoding, thereby reducing the deleterious effect of categorization on the late-selection process. However, no evidence for this

reduction was found. Instead, the late-selection process was impaired by categorization, regardless of the type of encoding, particularly in free recall.

As in Experiment 1, free-report quantity, the typical measure used in many studies of recall, showed a reduced effect of categorization relative to forced-report quantity; stated differently, the difference between free and forced report increased as the number of categories in the study list decreased (Fig. 2). We attribute this reduction to resolution having an impact on free-report quantity, which it does not have on forced-report quantity. In other words, the beneficial effects of categorization on the early-selection process were offset by impairment to the late-selection process, and this offsetting seemed to occur regardless of how the study pairs were encoded.

The analysis of the free- versus forced-report quantity difference also demonstrated that it was affected by encoding type, but only in the cued-recall group. That is, a larger quantity difference emerged if the study pairs were encoded with rote repetition rather than interactive imagery (cf. panels A and B in Fig. 2). This effect occurred despite the fact that resolution did not differ between the conditions (Table 2). How then, do we account for this difference?

As we discussed in the introduction, free-report quantity is not just a measure reflecting early- and late-selection influences of recall, but is affected by report bias as well. Examination of the type-2 HRs and FARs (the proportions of correct vs. incorrect responses, respectively, that were reported) suggests that the effect of encoding on the quantity difference in the cued-recall group was indeed attributable to bias. Specifically, for interactive-imagery encoding, both the HR ($M = .95$, $SEM = .01$) and the FAR ($M = .23$, $SEM = .05$) were greater than for rote-repetition encoding (HR, $M = .84$, $SEM = .04$; FAR, $M = .13$, $SEM = .03$): for the main effect of encoding, $F(1, 24) = 10.40$, $p = .004$, $\eta^2 = .30$. In contrast, in the free-recall group, the HRs and FARs were remarkably similar between the rote-repetition (HR, $M = .92$, $SEM = .02$; FAR, $M = .25$, $SEM = .05$) and interactive-imagery (HR, $M = .91$, $SEM = .03$; FAR, $M = .24$, $SEM = .05$) conditions: $F < 1$ for the main effect of encoding. Computing C , a common measure of bias in SDT, from these rates yields a large difference between the encoding conditions in the cued-recall group (rote repetition, $-.02$; interactive imagery, $-.61$) but very little difference in the free-recall group (rote repetition, $-.36$; interactive imagery, $-.33$). Thus, in the cued-recall group only, the more conservative criterion placement in the rote-repetition condition than in the interactive-imagery condition would have provided more of an opportunity to reveal correctly withheld responses, as responding was forced. Revealing more of these correct responses would have, in turn, led to the greater increase in quantity that was observed.

General discussion

A long line of memory research has shown that recall performance is enhanced by categorical organization of the study materials (e.g., Cofer et al., 1966; Guerin & Miller, 2008; Howard & Kahana, 2002; Puff et al., 1977; Runquist, 1970; Tehan, 2010; Thompson et al., 1972; Tulving & Pearlstone, 1966). In contrast, categorical organization generally impairs recognition memory performance (e.g., Dewhurst, 2001; Hintzman, 1988; Koutstaal & Schacter, 1997; Shiffrin et al., 1995), although some exceptions to this general rule have been found (e.g., Bruce & Fagan, 1970; Connor, 1977; D'Agostino, 1969; Kintsch, 1968; Mandler, 1972; Neely & Balota, 1981). On the basis of these findings, we predicted that early- and late-selection processes of cued and free recall would also dissociate, with the early-selection process showing a benefit from categorical structure and the late-selection process showing a deficit. To separate these processes, type-2 SDT methodology incorporating a report option and confidence ratings was used. Forced-report quantity and resolution were designated as the measures of early and late selection, respectively. As predicted, categorical organization of the targets in the study list facilitated forced-report quantity but impaired resolution in two separate experiments. Furthermore, although free-report quantity, the typical measure of recall in experiments in which participants are instructed not to guess, also showed facilitation from categorical structure, the facilitation was not as great as that for forced-report quantity. We argued that this was due to the late-selection process of recall offsetting the facilitation of free- but not of forced-report quantity.

The present study has added to a growing body of research that has focused on separating the early- and late-selection processes of memory. For example, Jacoby and colleagues (e.g., Jacoby, Kelley, & McElree, 1999; Jacoby, Shimizu, Daniels, & Rhodes, 2005a; Jacoby, Shimizu, Velanova, & Rhodes, 2005b; see also Alban & Kelley, 2012, and Marsh et al., 2009) have explored *source-constrained retrieval*, an early selection mechanism in memory. Source-constrained retrieval was initially investigated in recognition memory tasks using the *memory-for-foils* procedure. In an initial demonstration of the effect, Jacoby, Shimizu, Velanova, and Rhodes (2005b) had their participants encode targets with either a deep (pleasantness judgment) or shallow (vowel counting) orienting task, and later they administered a recognition memory test. Subsequently, the foils in the recognition memory test were intermixed with new foils, and recognition memory was tested again, with the expectation that the old foils from the first recognition memory test should be called “old,” whereas new foils introduced on the second recognition memory test should be called “new.” Source-constrained retrieval was shown to occur in that old foils that had followed a

deep orienting task were recognized better on the second recognition memory test than were the old foils that had followed a shallow one. In other words, the processing of *foils* on the first recognition memory test corresponded to the processing engaged in at study. Jacoby, Shimizu, Velanova, and Rhodes explained this effect by assuming that participants strategically implemented qualitatively different retrieval processes at test following a study phase with a deep versus a shallow orienting task. In other words, participants appeared to be strategically mentally reinstating the orienting task at test that had been used at study, in an attempt to aid retrieval. Note that this strategic choice of retrieval processing is an early- rather than a late-selection process.

More recently, Halamish et al. (2012; see also Wahlheim & Jacoby, 2011, Exp. 3) extended the memory-for-foils paradigm to investigate early-selection processes in cued recall. In their Experiment 2, they adopted a method similar to that used by Weldon and Colston (1995). In particular, they instructed participants during a forced-report stage of the experiment to write down all candidates that came to mind during the retrieval attempt, regardless of whether or not they believed them to be targets. Participants were to stop production only when they thought that they had retrieved the target (or if they could no longer produce any better candidates) and to mark the candidate that they believed was most likely the target. Halamish et al. then computed the percentage of first-candidate responses that were targets, which was their estimate of the early-selection process. Their measurement of first-candidate target percentage, rather than the percentage of targets ultimately recalled, was done in an attempt to limit late-selection contamination of the measure.

This procedure resembles similar ones that have emerged over the years in recall research in an attempt to isolate the early-selection process (see, e.g., Bousfield & Rosner, 1970; Guynn & McDaniel, 1999; Higham & Tam, 2005; Kahana, Dolan, Sauder, & Wingfield, 2005). For example, Higham and Tam (2005, Exp. 3) had participants produce up to six responses per cue in a cued-recall task. If the target was amongst any of the generated candidates, the trial was counted as correct. In our view, Higham and Tam's methodology, and ones like it for estimating the early-selection process (e.g., uninhibited report instructions; Bousfield & Rosner, 1970), is less likely to be contaminated by late-selection processes than is the one adopted by Halamish et al. (2012) because, with the latter procedure, participants may redefine the cue set during the process of retrieval. Such redefinition (e.g., considering alternative meanings of the cue word and interrogating memory with the new meaning) would allow a target to be accessed, say, in the third serial position purely because the early-selection process had changed, not because the late-selection process had contaminated it.

In the present research, we had participants only produce one candidate response in a forced report, and critics may argue that late-selection processes may affect the decision of which candidate to offer if more than one was covertly being considered for report. We agree that forced-report quantity involving only one candidate response likely incorporates some late-selection filtering. However, that forced-report quantity is not a *pure* measure of the early-selection process is not critical for the present purposes, as long as it is less contaminated than free report, which it necessarily must be. The primary goal here was to use the type-2 SDT methodology to separate enough of the late-selection process from the early one to determine whether categorization has opposing effects on recall. The results of these two experiments clearly indicate that such opposing influences do exist; it was not necessary to derive *process-pure* estimates of early- and late-selection processes to meet this end.

Resolution versus overt recognition

In both experiments, resolution was impaired by categorical structure. As noted above, this finding is similar to that found with overt recognition, a finding that is attributable to test lures that belong to the same category sharing features with multiple studied exemplars, causing the FAR to increase (e.g., Hintzman, 1988). Indeed, we found that the poor resolution for categorized lists was also partly attributable to a high FAR. For example, in Experiment 1, we found a significant increase in the FAR (the proportion of incorrect responses that were reported) as the number of categories decreased, $F(1, 116) = 3.12, p = .05, \eta^2 = .05$ (two-category, $M = .32, SEM = .04$; six-category, $M = .29, SEM = .04$; 24-category, $M = .24, SEM = .04$).

Recognition studies have also found that categorization causes the HR to increase slightly, presumably because of very high similarity between old test items and their representations in memory (e.g., Arndt & Hirshman, 1998). However, in contrast to this result with overt recognition, we found that the type-2 HR tended to *reduce* with fewer categories. For example, a significant main effect of category number on the HR (the proportion of correct responses that were reported) was found in Experiment 1, $F(1, 116) = 3.74, p = .03, \eta^2 = .06$ (two-category, $M = .89, SEM = .02$; six-category, $M = .87, SEM = .03$; 24-category, $M = .94, SEM = .02$).

The fact that categorization tended to decrease rather than increase the HR raises a potential criticism of the present work. As noted above, to compute resolution in recall, it is necessary to know the responses that participants withheld. Without knowing these responses, it is impossible to compute the type-2 HRs and FARs, which are required to estimate resolution. However, the instruction to respond to every cue in cued recall, or to provide a set number of

responses that correspond to the number of study items in free recall, may have implemented a generation strategy that would not normally occur if participants had been instructed not to guess. Thus, the critic could argue, the poor resolution with categorized lists is attributable to participants generating associates to the studied categories but opting to withhold them in order to meet the unusual demands of forced report. Because of the categorical structure, this strategy would result in a number of “lucky guesses,” or responses that were correct but withheld because participants did not believe them to be targets. Indeed, “metacognitive misses” of this sort must be occurring for the HR to reduce, and for the difference between the free- and forced-report quantities to increase (Figs. 1 and 2) with a reduction in the number of categories in the study list.

Our response to this potential criticism is twofold. First, we admit that some form of generation strategy was likely adopted by some of our participants some of the time, particularly if memory for targets was poor. However, in some sense, the fact that participants were able to generate targets using the categorical structure, but then fail to identify them as previously studied items (and, hence, fail to report them), is exactly the point; this shows that the categorical structure made discrimination between studied and nonstudied exemplars of the same category difficult. It is unlikely, for example, that had a target been considered for report following an uncategorized study list that the same recognition failure would have occurred. Our second response is that, although some lucky guessing may have occurred for the categorized lists, it was not the sole problem with resolution. As already noted above, an increased FAR for categorized lists was also a contributing factor.

Nonetheless, different methodologies might be adopted to limit a generation strategy in free report, ensuring that at least those responses are more akin to ones obtained in more typical recall experiments. For example, rather than demanding responses along with a free/forced decision during the first pass through the test, a two-stage procedure could be used whereby omissions are allowed during the first stage, but some form of response to all cues is required in a second stage toward the end of the experiment. As long as participants are not made aware that responding will be forced later in the experiment, this procedure would presumably discourage them from adopting a generate-recognize strategy during free report. Higham (2002) adopted such a procedure in the *phase* group of a cued-recall experiment and compared the results from that experimental group with those from a *trial* group, in which participants were required to make free/forced decisions on a trial-by-trial basis. He found that the results between the groups were highly comparable, despite the fact that the trial group were aware throughout the test phase that some form of response would ultimately be required to each cue, whereas

the phase group were not. These results suggest that a priori knowledge that a response is required on every trial is not enough to initiate a generate-recognize strategy in free report. Also, trial-by-trial free/forced decisions sidestep other potential problems introduced by obtaining output from participants in separate stages of the experiment. Because participants essentially have two temporally separated attempts at producing the target to a given cue, the separate-phase procedure introduces confounding variables such as the number of retrieval attempts and retrieval processing time, as well as allowing hypermnnesia. Overall, then, the trial-by-trial method for obtaining responses and free/forced decisions would seem to be the preferable methodology and may partly explain why it is being increasingly adopted in research examining strategic accuracy regulation (e.g., Higham, 2007, Exp. 2; Jacoby, Wahlheim, Rhodes, Daniels, & Rogers, 2010; Wahlheim & Jacoby, 2011).

Returning to the present results, the fact that categorization decreased rather than increased the HR marks a difference between our results and those obtained in overt-recognition studies. One potential cause of the difference is that studies that have demonstrated an increase to the HR from categorization have used old/new recognition. That is, participants are required to decide whether a *single item* is a studied or nonstudied exemplar from a categorized list. This differs from the type-2 report/withhold decision with confidence assignment, in that participants may have been considering more than one candidate at a time. An interesting avenue for future research would be to examine the effect of categorization on the ability to discriminate studied from nonstudied targets in a two-, three-, or four-alternative forced choice recognition design using a report option. The results from such a study, in which participants must discriminate between multiple exemplars on a given test trial, may show that categorization produces an increase in the FAR as well as a decrease the HR, just as was shown for resolution in the present experiments.

Report bias

Most of our discussion has focused on the decrease to free- as compared to forced-report quantity caused by categorization's impairment of the late-selection process. However, as noted in the introduction, the type-2 SDT framework also predicts that report bias—the position of the type-2 report criterion—can also influence free-report quantity. If this criterion is set conservatively, there is more opportunity for metacognitive misses (withheld correct responses) to increase quantity as responding is forced, relative to when the criterion is set liberally. Such an influence was demonstrated in the cued-recall group of Experiment 2. In particular, if participants were instructed to encode the study pairs

with rote repetition, they were unbiased ($C = -0.02$). However, for interactive-imagery encoding, the criterion was set liberally ($C = -0.61$). This liberal criterion-setting for interactive-imagery encoding resulted in a smaller difference between the free- and forced-report quantities than was present for rote-repetition encoding (see panels A and B in Fig. 2). Interestingly, report criterion differences were not observed in free recall. We attribute this pattern of results to participants' assessment of the difficulty of the test. Interactive-imagery encoding was particularly facilitative in the cued-recall group, a fact that may have encouraged participants to set their criterion liberally, maximizing the HR with little cost to the FAR.

Conclusions

Many memory studies over many years have reached the same conclusion: that categorical structure has a facilitative effect on recall. The mechanism by which this occurs is generally thought to be cueing. As stated by Guerin and Miller (2008), "It seems likely that [categorization] facilitates recall because it facilitates access to information during retrieval by providing a potent, easily accessible cue" (p. 302). Conceptualized in terms of early- versus late-selection processes, it is clear from this statement that categorization, by facilitating "access" to memory traces, operates via an early-selection process. By using the type-2 SDT framework to separate the early- and late-selection processes of recall, our research has demonstrated that the facilitative effect of categorization on the early-selection process is offset by a detrimental effect on the late one, suggesting that previous research has underestimated categorization's true cueing effect. We encourage researchers interested in investigating the effects of categorization and other organizational aids on recall to employ the type-2 SDT framework in the future so that early-selection processes, late-selection processes, and response strategies (i.e., criterion setting) can be teased apart.

Author note This research was part of M.A.G.'s PhD thesis and was financially supported by ORSAS (Overseas Research Students Awards Scheme) and by the School of Psychology at the University of Southampton, U.K. Portions of this research were presented at the 1st Joint EPS (Experimental Psychology Society)/SEPEX (Spanish Society of Experimental Psychology) Conference on April 17, 2010, in Granada, Spain, and at the 51st Annual Meeting of the Psychonomic Society on November 19, 2010, in St. Louis, Missouri, USA.

References

- Alban, M. W., & Kelley, C. M. (2012). Variations in constrained retrieval. *Memory & Cognition*, 40, 681–692. doi:10.3758/s13421-012-0185-5

- Arndt, J., & Hirshman, E. (1998). True and false recognition in MINERVA2: Explanations of a global matching perspective. *Journal of Memory and Language*, 39, 371–391. doi:10.1006/jmla.1998.2581
- Bousfield, W. A., & Rosner, S. R. (1970). Free versus uninhibited recall. *Psychonomic Science*, 20, 75–76.
- Bower, G. H. (1970). Imagery as a relational organizer in associative learning. *Journal of Verbal Learning and Verbal Behavior*, 9, 529–533. doi:10.1016/S0022-5371(70)80096-2
- Bower, G. H., & Winzenz, D. (1970). Comparison of associative learning strategies. *Psychonomic Science*, 20, 119–120.
- Bruce, D., & Fagan, R. L. (1970). More on the recognition and free recall of organized lists. *Journal of Experimental Psychology*, 85, 153–154. doi:10.1037/h0029532
- Cofer, C. N., Bruce, D. R., & Reicher, G. M. (1966). Clustering in free recall as a function of certain methodological variables. *Journal of Experimental Psychology*, 71, 858–866. doi:10.1037/h0023217
- 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 and Verbal Behavior*, 11, 671–684. doi:10.1016/S0022-5371(72)80001-X
- D'Agostino, P. R. (1969). The blocked-random effect in recall and recognition. *Journal of Verbal Learning and Verbal Behavior*, 8, 815–820. doi:10.1016/S0022-5371(69)80048-4
- Dewhurst, S. A. (2001). Category repetition and false recognition: Effects of instance frequency and category size. *Journal of Memory and Language*, 44, 153–167. doi:10.1006/jmla.2000.2738
- Donaldson, W. (1992). Measuring recognition memory. *Journal of Experimental Psychology: General*, 121, 275–277. doi:10.1037/0096-3445.121.3.275
- Dunlosky, J., & Hertzog, C. (1998). Aging and deficits in associative memory: What is the role of strategy use? *Psychology and Aging*, 13, 597–607. doi:10.1037/0882-7974.13.4.597
- Goldsmith, M. (2011). Quantity-accuracy profiles or type-2 signal detection measures? Similar methods toward a common goal. In P. A. Higham & J. P. Leboe (Eds.), *Constructions of remembering and metacognition: Essays in honour of Bruce Whittlesea* (pp. 128–136). Basingstoke, U.K.: Palgrave-Macmillan.
- Goldsmith, M., & Koriat, A. (2008). The strategic regulation of memory accuracy and informativeness. In A. S. Benjamin & B. H. Ross (Eds.), *The psychology of learning and motivation* (Memory use as skilled cognition, Vol. 48, pp. 1–60). San Diego, CA: Elsevier Academic Press.
- Green, D. M., & Moses, F. L. (1966). On the equivalence of two recognition measures of short-term memory. *Psychological Bulletin*, 66, 228–234. doi:10.1037/h0023645
- Guerin, S. A., & Miller, M. B. (2008). Semantic organization of study materials has opposite effects on recognition and recall. *Psychonomic Bulletin & Review*, 15, 302–308. doi:10.3758/PBR.15.2.302
- Guynn, M. J., & McDaniel, M. A. (1999). Generate—Sometimes recognize, sometimes not. *Journal of Memory and Language*, 41, 398–415. doi:10.1006/jmla.1999.2652
- Halamish, V., Goldsmith, M., & Jacoby, L. L. (2012). Source-constrained recall: Front-end and back-end control of retrieval quality. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 38, 1–15. doi:10.1037/a0025053
- Higham, P. A. (2002). Strong cues are not necessarily weak: Thomson and Tulving (1970) and the encoding specificity principle revisited. *Memory & Cognition*, 30, 67–80. doi:10.3758/BF03195266
- Higham, P. A. (2007). No Special K! A signal detection framework for the strategic regulation of memory accuracy. *Journal of Experimental Psychology: General*, 136, 1–22. doi:10.1037/0096-3445.136.1.1
- Higham, P. A. (2011). Accuracy discrimination and type-2 signal detection theory: Clarifications, extensions, and an analysis of bias. In P. A. Higham & J. P. Leboe (Eds.), *Constructions of remembering and metacognition: Essays in honour of Bruce Whittlesea* (pp. 109–127). Basingstoke, U.K.: Palgrave Macmillan.
- Higham, P. A., & Arnold, M. M. (2007). How many questions should I answer? Using bias profiles to estimate optimal bias and maximum score on formula-scored tests. *European Journal of Cognitive Psychology*, 19, 718–742. doi:10.1080/09541440701326121
- Higham, P. A., & Tam, H. (2005). Generation failure: Estimating metacognition in cued recall. *Journal of Memory and Language*, 52, 595–617. doi:10.1016/j.jml.2005.01.015
- Hintzman, D. L. (1988). Judgments of frequency and recognition memory in a multiple-trace memory model. *Psychological Review*, 95, 528–551. doi:10.1037/0033-295X.95.4.528
- Howard, M. W., & Kahana, M. J. (2002). When does semantic similarity help episodic retrieval? *Journal of Memory and Language*, 46, 85–98. doi:10.1006/jmla.2001.2798
- Jacoby, L. L., Kelley, C. M., & McElree, B. D. (1999). The role of cognitive control: Early selection versus late correction. In S. Chaiken & Y. Trope (Eds.), *Dual-process theories in social psychology* (pp. 383–400). New York, NY: Guilford Press.
- Jacoby, L. L., Shimizu, Y., Daniels, K. A., & Rhodes, M. G. (2005a). Modes of cognitive control in recognition and source memory: Depth of retrieval. *Psychonomic Bulletin & Review*, 12, 852–857. doi:10.3758/BF03196776
- Jacoby, L. L., Shimizu, Y., Velanova, K., & Rhodes, M. G. (2005b). Age differences in depth of retrieval: Memory for foils. *Journal of Memory and Language*, 52, 493–504. doi:10.1016/j.jml.2005.01.007
- Jacoby, L. L., Wahlheim, C. N., Rhodes, M. G., Daniels, K. A., & Rogers, C. S. (2010). Learning to diminish the effects of proactive interference: Reducing false memory for young and older adults. *Memory & Cognition*, 38, 820–829. doi:10.3758/MC.38.6.820
- Kahana, M. J., Dolan, E. D., Sauder, C. L., & Wingfield, A. (2005). Intrusions in episodic recall: Age differences in editing of overt responses. *Journals of Gerontology*, 60B, P92–P97. doi:10.1093/geronb/60.2.P92
- Kintsch, W. (1968). Recognition and free recall of organized lists. *Journal of Experimental Psychology*, 78, 481–487. doi:10.1037/h0026462
- Koriat, A., & Goldsmith, M. (1996). Monitoring and control processes in the strategic regulation of memory accuracy. *Psychological Review*, 103, 490–517. doi:10.1037/0033-295X.103.3.490
- Koriat, A., Goldsmith, M., & Halamish, V. (2008). Controlled processes in voluntary remembering. In J. Byrne & H. L. Roediger III (Eds.), *Learning and memory: A comprehensive reference*. Vol. 2: *Cognitive psychology of memory* (pp. 307–324). Oxford, U.K.: Elsevier.
- Koutstaal, W., & Schacter, D. L. (1997). Gist-based false recognition of pictures in older and younger adults. *Journal of Memory and Language*, 37, 555–583. doi:10.1006/jmla.1997.2529
- Kučera, H., & Francis, W. N. (1967). *Computational analysis of present-day American English*. Providence, RI: Brown University Press. Retrieved from www.psych.rl.ac.uk/MRC_Psych_Db_files/mrc2.html
- Lueddeke, S., & Higham, P. A. (2011). Expertise and gambling: Using type-2 signal detection theory to investigate differences between regular gamblers and non-gamblers. *Quarterly Journal of Experimental Psychology*, 64, 1850–1871. doi:10.1080/17470218.2011.584631
- Luna, K., Higham, P. A., & Martin-Luengo, B. (2011). Regulation of memory accuracy with multiple answers: The plurality option.

- Journal of Experimental Psychology: Applied*, 17, 148–158. doi:10.1037/a0023276
- Macmillan, N. A., & Creelman, C. D. (2005). *Detection theory: A user's guide* (2nd ed.). New York, NY: Erlbaum.
- Mandler, G. (1972). Organization and recognition. In E. Tulving & W. Donaldson (Eds.), *Organization and memory* (pp. 139–166). New York, NY: Academic Press.
- Marsh, R. L., Meeks, J. T., Cook, G. I., Clark-Foos, A., Hicks, J. L., & Brewer, G. A. (2009). Retrieval constraints on the front end create differences in recollection on a subsequent test. *Journal of Memory and Language*, 61, 470–479. doi:10.1016/j.jml.2009.06.005
- Masson, M. E. J., & Rotello, C. M. (2009). Sources of bias in the Goodman-Kruskal gamma coefficient measure of association: Implications for studies of metacognitive processes. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 35, 509–527. doi:10.1037/a0014876
- 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
- Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (1998). The University of South Florida word association, rhyme, and word fragment norms. Retrieved from <http://w3.usf.edu/FreeAssociation/>
- Pollack, I., Norman, D. A., & Galanter, E. (1964). An efficient non-parametric analysis of recognition memory. *Psychonomic Science*, 1, 327–328.
- Puff, C. R. (1970). Role of clustering in free recall. *Journal of Experimental Psychology*, 86, 384–386. doi:10.1037/h0030189
- Puff, C. R., Murphy, M. D., & Ferrara, R. A. (1977). Further evidence about the role of clustering in free recall. *Journal of Experimental Psychology: Human, Learning, and Memory*, 3, 742–753. doi:10.1037/0278-7393.3.6.742
- Robbins, D., Bray, J. F., Irvin, J. R., & Wise, P. S. (1974). Memorial strategy and imagery: An interaction between instructions and rated imagery. *Journal of Experimental Psychology*, 102, 706–709. doi:10.1037/h0036116
- Rotello, C. M., Masson, M. E. J., & Verde, M. F. (2008). Type I error rates and power analysis for single-point sensitivity measures. *Perception & Psychophysics*, 70, 389–401. doi:10.3758/PP.70.2.389
- Runquist, P. A. (1970). Clustering in free recall following paired associate learning. *Journal of Experimental Psychology*, 83, 260–265. doi:10.1037/h0028546
- Shiffrin, R. M., Huber, D. E., & Marinelli, K. (1995). Effects of category length and strength on familiarity in recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 21, 267–287. doi:10.1037/0278-7393.21.2.267
- Tehan, G. (2010). Associative relatedness enhances recall and produces false memories in immediate serial recall. *Canadian Journal of Experimental Psychology*, 64, 266–272. doi:10.1037/a0021375
- Thompson, C. P., Hamlin, V. J., & Roenker, D. L. (1972). A comment on the role of clustering in free recall. *Journal of Experimental Psychology*, 94, 108–109. doi:10.1037/h0032797
- Tulving, E., & Pearlstone, Z. (1966). Availability versus accessibility of information in memory for words. *Journal of Verbal Learning and Verbal Behavior*, 5, 381–391. doi:10.1016/S0022-5371(66)80048-8
- Van Overschelde, J. P., Rawson, K. A., & Dunlosky, J. (2004). Category norms: An updated and expanded version of the Battig and Montague (1969) norms. *Journal of Memory and Language*, 50, 289–335. doi:10.1016/j.jml.2003.10.003
- Wahlheim, C. N., & Jacoby, L. L. (2011). Experience with proactive interference diminishes its effects: Mechanisms of change. *Memory & Cognition*, 39, 185–195. doi:10.3758/s13421-010-0017-4
- Weldon, M. S., & Colston, H. L. (1995). Dissociating the generation stage in implicit and explicit memory tests: Incidental production can differ from strategic access. *Psychonomic Bulletin & Review*, 2, 381–386. doi:10.3758/BF03210976