Late consequences of early selection: When memory monitoring backfires

Katarzyna Zawadzka, Maciej Hanczakowski, and Edward L. Wilding

Cardiff University, UK

Word count: 10,058

Author note

Katarzyna Zawadzka, Maciej Hanczakowski, and Edward L. Wilding, School of Psychology, Cardiff University, UK.

Katarzyna Zawadzka is now at Division of Psychology, Nottingham Trent University.

Edward L. Wilding is now at School of Psychology, University of Nottingham.

Correspondence concerning this article should be addressed to Katarzyna Zawadzka, Division of Psychology, Nottingham Trent University, Burton Street, Nottingham NG1 4BU, UK. Email: katarzyna.zawadzka@ntu.ac.uk

Article published in Journal of Memory and Language. DOI: 10.1016/j.jml.2016.06.003

Abstract

At retrieval, people can adopt a retrieval orientation by which they recreate the mental operations used at encoding. Monitoring by retrieval orientation leads to assessing all test items for qualities related to the encoding task, which enriches foils with some of the qualities already possessed by targets. We investigated the consequences of adopting a retrieval orientation under conditions of repeated monitoring of the same foils. Participants first processed foils in the context of one of two tests encouraging different retrieval orientations. The foils were then re-used on a subsequent test in which retrieval orientation either matched or mismatched that adopted on the first test. In the aggregate data, false alarms for repeated foils were higher when there was a match between the retrieval orientations on both tests. This demonstrates that when retrieval orientation enriches foils with target-like characteristics, it can backfire when repeated monitoring of the same foils is required.

Keywords: retrieval orientation, monitoring, memory for foils

Late consequences of early selection: When memory monitoring backfires

When asked to retrieve some information from memory, people can employ a variety of monitoring strategies to improve the quality of their memory report. One strategy is to mentally recreate the operations performed at the time of encoding. This mentally recreated information becomes embedded in the retrieval cue and every item in a memory test is then assessed with respect to the degree of match with this cue. Since only studied items are associated with diagnostic details now embedded in the retrieval cue, this form of monitoring allows for effective rejection of non-studied items (foils). This monitoring strategy can be viewed as a consequence of having adopted a *retrieval orientation* (e.g., Gray & Gallo, 2015; Herron & Rugg, 2003a; Pierce & Gallo, 2011; Rugg & Wilding, 2000) or as an example of *early selection* (e.g., Guzel & Higham, 2013; Jacoby, Kelley, & McElree, 1999).[1]

Research conducted to date has shown the benefits of applying early selection mechanisms for memory reports (e.g., Bridger, Herron, Elward, & Wilding, 2009; Pierce & Gallo, 2011). The present study breaks with this tradition by delineating the conditions in which the use of such a monitoring strategy comes at a cost when repeated monitoring of the same foils is required.

Evidence for monitoring by retrieval orientation comes from two strands of research which, although distinct, share a common approach: they infer the

[1] Other terms that have been used to describe this kind of monitoring process include *front-end control* (e.g., Halamish, Goldsmith, & Jacoby, 2012), and *source-constrained retrieval* (e.g., Alban & Kelley, 2012; Jacoby, Shimizu, Daniels, & Rhodes, 2005; Jacoby, Shimizu, Velanova, & Rhodes, 2005). Generally, early selection monitoring strategies are contrasted with strategies referred to as *late correction* (Jacoby et al., 1999), such as the distinctiveness heuristic (Dodson & Schacter, 2002; Hanczakowski & Mazzoni, 2011) or response withholding (Koriat & Goldsmith, 1996), which operate on the information already retrieved from memory.

operation of a monitoring strategy from the ways in which foils are processed at test. Both strands capitalize on the premise that as non-studied foils are by definition unaffected by the study phase, any difference between the processing of the foils must be caused by the monitoring strategy adopted at test. The first strand of research uses measures of neural activity such as event-related potentials (ERPs). It is possible to compare ERPs elicited by foils in two tasks differing with respect to the kind of information that needs to be retrieved in order to answer a memory question. The differences between ERPs are assumed to demonstrate the operation of distinct retrieval orientations depending on the type of queried information. Past research has shown differences between ERPs elicited by new items when study items were presented as pictures versus words (e.g., Herron & Rugg, 2003a, Robb & Rugg, 2002), were studied with a pleasantness versus an animacy judgment (Herron & Wilding, 2004, 2006), or with a shallow versus deep processing task (Rugg, Allan, & Birch, 2000), among others. These studies clearly demonstrate that non-studied foils are processed differently under various retrieval orientations.

The second strand of research uses behavioral methods to gain insights into when and how monitoring by retrieval orientation is employed. In the *memory-for-foils paradigm* (Jacoby, Shimizu, Daniels, & Rhodes, 2005; Jacoby, Shimizu, Velanova, & Rhodes, 2005; Shimizu & Jacoby, 2005), participants first learn a list of words with two different orienting tasks: one deep (for example, a pleasantness judgment task) and one shallow (e.g., counting the number of letters or vowels in each studied word). Following the study phase, they are given two old/new recognition tests. On the deep test, only deeply processed

words are presented among foils. On the shallow test, only words studied with the shallow task are among foils. Finally, an additional memory test for unstudied foils is administered. The final test list consists of three types of items: deep foils (foils presented on the deep test), shallow foils (foils presented on the shallow test), and new words (not presented on any of the tests). Participants are instructed to distinguish new words from those that were presented earlier during any phase of the experiment. The main finding in this paradigm is a task-dependent difference in correct endorsements for previously encountered foils: deep foils are more often indicated as having been seen during the course of the experiment than shallow foils. This is taken as evidence that the retrieval orientation adopted on the deep test benefits the subsequent memorability of the foils on that task to a greater degree than the orientation adopted on the shallow test. Crucially, this line of research goes beyond demonstrating that non-studied foils are subjected to different assessments. The novel focus here is on the *consequences* of adopting a retrieval orientation: Jacoby, Shimizu, Daniels, and Rhodes (2005) argued that the assessment of foils with the use of retrieval orientation on the deep test enriches these foils with diagnostic information embedded in the retrieval cue.

The findings in subsequent studies support the explanation that the better memory for deep foils observed by Jacoby, Shimizu, Daniels, and Rhodes (2005) stems from their enrichment with details diagnostic of study. Marsh et al. (2009; see also Danckert, MacLeod, & Fernandes, 2011; Gray & Gallo, 2015) added a remember/know task (e.g., Tulving, 1985; Gardiner, 1988) to the final test for foils. Foils presented on a deep test were later assigned ‘remember’

responses more often than foils first encountered on a shallow test. This is consistent with the assumption that on deep tests new words are deeply processed in the context of the orienting task. Danckert et al. (2011) substituted the final test for foils with a second judgment phase, in which the task was to perform on foils the same judgments that were earlier made for targets at study. They predicted that if participants processed foils in the context of a retrieval orientation on the deep test, then a judgment had already been made for deep foils when they were first presented. If this was the case, then a deep orienting task during the second judgment phase would be completed faster for deep than for shallow foils. The results were consistent with that prediction. Recently, Gray and Gallo (2015) demonstrated that the deep > shallow difference in memory for foils occurs at all levels of foil strength, ruling out an alternative explanation that this effect is due to a post-retrieval monitoring process employed specifically for items yielding ambiguous evidence that does not allow a determination of whether an item was studied or not.

The research conducted to date allows a clear conclusion: monitoring by early selection can change the way foils are processed in a memory test. Specifically, foils are considered in light of the adopted retrieval orientation which, if the test requires deep processing, leads to their enrichment with the details that are embedded in the particular retrieval cue. In other words, deeply processed foils become associated with the details that are diagnostic of previous study. If a memory test for foils is later given, such enriched foils are remembered better than foils that were monitored on a shallow test.

However, if monitoring by retrieval orientation can enrich foils with features diagnostic of earlier study, it means that foils may start to resemble studied items. A straightforward question thus arises: what would be the consequences of early selection if the same retrieval orientation was applied twice to the same foils? If applying a retrieval orientation makes foils more similar to targets, would people mistake these foils for targets on a subsequent test requiring the adoption of the same retrieval orientation? In other words, could adopting a potentially beneficial monitoring strategy at test ultimately lead to impairment in performance on a future test if the same to-be-rejected materials are assessed again?

We tested this hypothesis in four experiments by having participants complete two study-test blocks. The first block consisted of a single study phase followed by two test phases. In the study phase, single words were studied in two different deep encoding tasks. The test phases were both exclusion tasks. Each required endorsement of words studied in one of the two encoding tasks, and rejection of new words as well as those studied in the alternate encoding task. The encoding task associated with words requiring endorsement was changed across the two exclusion test phases.

The second study-test block had the same structure. The study phase was as for the first block, and all of the words presented were new to the experiment. The two exclusion tests also had the same structure, and the class of words requiring endorsement again varied across test phases. The critical manipulation was for the words that were to be rejected in each exclusion task in the second block. New words encountered in the exclusion tasks in the first block were re-

used as foils in the exclusion tasks in the second block. These foils were separated so that within each of the exclusion tasks there was an equal number that had been encountered as foils in each of the two exclusion tasks in the first block. This meant that there were foils for which the task demands matched across the two test phases in which they were encountered, and foils for which there was a mismatch.

Requiring participants to search for a particular set of items on the exclusion tests was aimed at encouraging them to adopt a specific retrieval orientation. Our measure of interest was the rate of foil endorsements on tests in the second block. We predicted that if applying a retrieval orientation to foils on the first test enriches them with qualitative characteristics that were searched for in targets, then on the second test those foils should be mistaken for targets more often if there was a match between the retrieval orientations adopted on the first and second test (*matched* condition) than if the retrieval orientations differed between the first and the second test (*mismatched* condition). In Experiment 1, we introduced this novel two-block procedure for investigating the aftereffects of foil processing. In Experiment 2, we increased the difficulty of the second-block tests by introducing a delay between the study and test phases in that block in order to increase the demand for monitoring at test. In Experiment 3, we changed one of the encoding tasks to establish the generalizability of the results obtained in Experiments 1 and 2. Finally, in Experiment 4, we modified the procedure with a view to constraining competing interpretations of our findings.

Experiment 1

Method

Participants. Twenty-four undergraduate students of Cardiff University (22 females; 18-25 years of age, *M* = 21.04, *SD* = 1.73) participated in this experiment for course credit or monetary compensation.

Materials and design. A total of 432 English singular nouns were chosen from the MRC database (Coltheart, 1981). Word length ranged from five to eight letters, and all words were of medium-to-high lexical frequency. In the first study-test block, 288 words were used. Half of these words served as study words, with 72 words studied with the pleasantness orienting question, requiring participants to provide a pleasantness judgment for each word, and 72 studied with the drawing orienting question, requiring participants to judge the ease of drawing a given item. The other half were introduced at test as non-studied foils. In the second study-test block, the remaining 144 words constituted the study list, with 72 words studied with each orienting task. At test, these studied words were accompanied by 144 words which had served as non-studied foils in the test phases in the first block. As the experiment was conducted in a within-participants design, all participants were presented with the full list of stimuli. The assignment of words to orienting tasks and to studied words versus foils was counterbalanced across participants.

Procedure. A schematic depiction of the procedure is presented in Figure 1. The first study-test block began with a self-paced study phase. Participants were presented with a list of single words, with each word accompanied by one of two orienting questions: a '*pleasantness'* question ("How pleasant is it?") or a '*drawing'* question ("How easy would it be to draw it?"). They were instructed

that their task was to answer the orienting question for each word on a 1-4 scale ('very unpleasant' to 'very pleasant' for the pleasantness task, and 'very difficult' to 'very easy' for the drawing task), as well as memorize both the word and the question for a future test. The 'pleasantness' and 'drawing' study trials were randomly intermixed within the study list. To facilitate distinguishing between the two orienting questions, the pleasantness question was always presented in purple font, and the drawing question in red font.

After the study phase, participants completed two separate exclusion tests (see e.g., Jacoby, 1991), one for each orienting task. On each test, participants were presented with 36 words studied with the 'pleasantness' question, 36 words studied with the 'drawing' question, and 72 non-studied foils. On a 'pleasantness' test (henceforth referred to as *P-test*), participants were asked to endorse as targets only those words that were earlier studied with the 'pleasantness' question. All other words, including those studied with the 'drawing' question, were to be rejected. On a 'drawing' test (*D-test*), only words studied with the 'drawing' question were to be endorsed as targets. The cue reminding participants which test they were completing (i.e. whether they should endorse words studied with the 'pleasantness' or 'drawing' question) was presented in the same color in which the orienting question was presented at study. The order of tests was counterbalanced across participants and fixed between the study-test blocks.

The second study-test block started with a study phase for a new list of words. Half of the list was studied with the 'pleasantness' question, and the other half with the 'drawing' question. The test phase began immediately after the

study phase. As in the first study-test block, participants took two exclusion tests, one for each orienting task. This time, however, the composition of the test lists differed from that from the first block. On each test, participants were presented with 36 words studied with the 'pleasantness' question, 36 words studied with the 'drawing' question, as well as with 72 words that were used in the first study-test block as non-studied foils. Thirty-six of these foils were presented in the first block on the P-test (henceforth referred to as *P-foils*), and 36 were presented in the first block on the D-test (*D-foils*). The instructions for participants were the same as in the first block: only words studied with the 'pleasantness' question in the second block were to be endorsed as targets on the P-test, and only words studied with the 'drawing' question were to be endorsed as targets on the D-test.

Results

Accuracy. We first compared accuracy (*d'*) on P- and D-tests across blocks (the descriptive statistics are presented in Table 1; see also Table 2 for raw hit and false alarm rates across experiments). For the sake of comparison between the first and second study-test blocks, the *d'* calculations were based only on endorsement rates for actually studied words (targets and foils studied with the other orienting task). The remaining non-studied foils were excluded from these calculations, as their status differed between the blocks (in the first block, they were new words, and in the second block the same foils were repeated).[2] A 2 (block: first, second) x 2 (test: P-test, D-test) repeated-measures Analysis of Variance (ANOVA) revealed only a significant main effect of block,

[2] For completeness, in the Appendix we present *d'* analyses for the full data set, including non-studied foils, for Experiments 1-4 (see also Table 1 for the descriptive statistics). In all experiments, the pattern of results was the same for the restricted and full data sets.

$F(1, 23) = 7.252$, $MSE = .18$, $p = .013$, $\eta_p^2 = .240$, which was driven by overall higher accuracy in the first ($M = 2.41$, $SD = 0.94$) than in the second block ($M = 2.18$, $SD = 1.09$). This could be explained by the fact that in the second block, test difficulty increased as compared to the first block as all words presented at test – including non-studied foils – were familiar. Also fatigue effects could have contributed to the lower second block accuracy. Crucially, the main effect of test was not significant, $F(1, 23) = 3.251$, $MSE = .39$, $p = .085$, $\eta_p^2 = .124$, showing that participants' accuracy was independent of the type of items searched for at test. The interaction was also not significant, $F < 1$. Together, these outcomes suggest that no differences according to foil type reported below can be attributed to differences in accuracy across tests.

Foil endorsement rates. Our main interest in this study lies in block-2 endorsement rates for foils first presented in block 1 (see the top-left panel of Figure 2 for the results). We predicted that matched foils, which had been presented twice on the same type of test, would be incorrectly endorsed as targets more often than mismatched foils, which were presented on two different tests. The data were consistent with this prediction, $t(23) = 2.104$, $SE = .01$, $p = .046$, $d = 0.23$. To further investigate this difference, we split the data depending on the type of test on which foils were presented for the first and second time (see the top-left panel of Figure 3). A 2 (foil type: P-foil, D-foil) x 2 (test 2 type: P-test, D-test) repeated-measures ANOVA revealed a significant foil x test interaction, $F(1,23) = 4.515$, $MSE = .002$, $p = .045$, $\eta_p^2 = .164$. Follow-up t-tests showed that P-foils were mistaken for targets more often when presented on a P-test than on a D-test, $t(23) = 2.752$, $SE = .01$, $p = .011$, $d = 0.36$. For D-foils,

however, the analogous difference was not significant, $t < 1$. The main effects of foil and test were not significant, $F < 1$ and $F(1,23) = 1.010$, $MSE = .005$, $p = .33$, $\eta_p^2 = .042$, respectively.

Discussion

Overall, the results confirmed the initial hypothesis: repeated monitoring of non-studied foils under the same retrieval orientation increased false alarms compared to when different retrieval orientations were encouraged. The matched > mismatched difference between foil endorsement rates that was present in the data collapsed across the tasks was significant, although small. Unexpectedly, this pattern was driven primarily by the differences between the endorsement rates of P-foils; the predicted pattern was not reliable for D-foils, even though numerically the trend was in the predicted direction. This asymmetry in endorsement rates for P- and D-foils was surprising, since we had no *a priori* reason for different patterns of results depending on which deep processing task was used.

Another issue with the present results is a generally low level of false alarms observed for repeated foils. The interpretation of any differences in false alarm rates is weakened when variability of these rates is truncated due to floor effects. False alarm rates are commonly low in recognition studies unless specific procedures are used to render foils very similar to studied items (e.g., Benjamin & Bawa, 2004; Roediger & McDermott, 1995). Given that we were interested in monitoring of foils that only after processing via retrieval orientation would become confusable with targets, low false alarm rates seem unavoidable in our design. Further experiments, described next, attest to the robustness of the

matched > mismatched pattern described here but the reader should interpret them in light of the fact that these pertain only to false alarm rates at the lower end of a scale.

As the results of Experiment 1 are, to our knowledge, the first demonstration of the costs of repeated monitoring of foils with the same retrieval orientation, in Experiment 2 we attempted to replicate this finding. We made only one change to the procedure used in Experiment 1. To increase the demand for monitoring on the final tests, in the second block we introduced a brief delay (see Figure 1 for the graphical representation). This delay was inserted between the study phase and the tests and lasted approximately 15 minutes, during which participants completed an unrelated cognitive task. Alban and Kelley (2012) have previously shown that participants are more likely to adopt a retrieval orientation as a means of monitoring foils in a recognition test if they expect this test to be difficult rather than easy. We assumed that separating the study and test phases would lead to increased difficulty of the second test, thus augmenting the chances of monitoring by retrieval orientation. We reasoned that increased application of retrieval orientation could reveal after-effects of task-specific processing of foils for both types of tasks rather than for one task only, as in Experiment 1.

Experiment 2

Method

Participants. Twenty-four undergraduate students of Cardiff University (22 females; 18-25 years of age, M = 19.92, SD = 1.77) participated in this experiment for course credit or monetary compensation.

Materials and procedure. The materials and design were the same as in Experiment 1. The procedure followed that from Experiment 1, with one exception. After the study phase, and before the test phase in the second study-test block, participants completed an unrelated cognitive task that lasted approximately 15 minutes.

Results

Accuracy. The descriptive statistics for *d'* are presented in Table 1. A 2 (block: first, second) x 2 (test: P-test, D-test) repeated-measures ANOVA performed on *d'* scores for studied words revealed only a significant main effect of block, $F(1, 23) = 26.028$, $MSE = .30$, $p < .001$, $\eta_p^2 = .531$, with higher accuracy in the first (M = 2.42, SD = 0.85) than in the second block (M = 1.85, SD = 0.89). The main effect of test was not significant, $F(1, 23) = 1.271$, $MSE = .27$, $p = .27$, $\eta_p^2 = .052$, and neither was the interaction, $F(1, 23) = 2.288$, $MSE = .24$, $p = .14$, $\eta_p^2 = .090$.

Foil endorsement rates. The results, presented in the bottom-left panel of Figure 2, replicated those from Experiment 1. Overall, matched foils were endorsed as targets more often than mismatched foils, $t(23) = 3.019$, $SE = .01$, $p = .006$, $d = 0.44$. The same results split by foil and test type are presented in the bottom-left panel of Figure 3. The 2 (foil type: P-foil, D-foil) x 2 (test 2 type: P-test, D-test) repeated-measures ANOVA revealed no effect of foil or test (both *F*s

< 1); the interaction, however, was again significant, $F(1, 23) = 9.293$, $MSE = .004$, $p = .006$, $\eta_p^2 = .288$. P-foils were mistaken for targets more often when presented on a P-test than on a D-test, $t(23) = 2.229$, $SE = .02$, $p = .036$, $d = 0.48$. Endorsement rates for D-foils did not differ significantly across tests, $t(23) = 1.384$, $SE = .02$, $p = .18$, $d = 0.28$.

Combined analyses. To investigate whether the lack of effect for D-foils was not due to low power, we binned the D-foil data from Experiments 1 and 2 (resulting in $N = 48$) and ran a mixed ANOVA on D- and P-test endorsement rates, with experiment as a between-subjects factor. Even in this case test type did not affect endorsement rates for D-foils, $F(1, 46) = 1.477$, $MSE = .006$, $p = .23$, $\eta_p^2 = .031$, although numerically the difference was in the predicted direction ($M = .11$, $SD = .12$ for endorsement rates on the D-test and $M = .09$, $SD = .09$ on the P-test). The main effect of experiment was not significant, and neither was the interaction, both Fs < 1.

In addition to that, we used the combined data from Experiments 1 and 2 to test whether the matched > mismatched pattern in the overall data set was consistent across participants, or whether it was driven by a small subset of participants whose results disproportionately contributed to the overall pattern. A two-tailed sign test revealed that the matched > mismatched pattern was indeed pervasive, $z = 2.499$, $p = .012$ (see Table 3 for the frequencies)[3].

Discussion

[3] If the data are analyzed separately for each of the experiments with two-tailed sign tests, $p = .078$ was obtained for Experiment 1, and $p = .115$ for Experiment 2.

Experiment 2 fully replicated the pattern of results in Experiment 1. Again, we found an overall increase in false alarms to foils subjected twice to monitoring with the same retrieval orientation in comparison to those subjected to different orientations. The brief delay between study and tests, which was aimed at increasing the difficulty of the block-2 tests, increased numerically the critical false alarm rates in comparison to those observed in Experiment 1. Although numerically the difference in endorsement rates for matched and mismatched foils was still small (~ 4 percentage points), in relative terms this amounts to almost half of the endorsement rate for mismatched foils.

Unexpectedly, the absence of a significant difference in D-foil endorsement rates was replicated as well, even in combined data from Experiments 1 and 2. There are at least two explanations for this result. First, it might be that the pleasantness test is unique in enriching foils with characteristics that can be retrieved on a subsequent test. In the memory-for-foils literature, the pleasantness task has been most often used as the deep orienting task so the majority of evidence for the consequences for foils of adopting retrieval orientation has come from this particular task. Still, there are studies that investigated the memory-for-foils phenomenon with other tasks requiring deep processing of targets (Alban & Kelley, 2012; Danckert et al., 2011) and given comparable results obtained in those studies, the idea of the uniqueness of the pleasantness task is unlikely to be correct.

Alternatively, it might be that the drawing task was not suited for the purpose of this experiment. Although it has been previously used in electrophysiological research to detect the operation of a retrieval orientation at

test (e.g., Bridger et al., 2009; Dzulkifli & Wilding, 2005), it is possible that the drawing task differs from other deep processing tasks in some aspect(s), for example due to its strong perceptual component. To assess this hypothesis of uniqueness of the drawing task, in Experiment 3 we substituted it with another deep processing task. Namely, we asked participants how often they could encounter the referent of each word in Cardiff (which was the city where our participants studied). In Experiment 3, we expected to document the pattern of increased false alarms to foils subjected twice to the same type of monitoring for both the pleasantness and the Cardiff task.

Experiment 3

Method

Participants. Forty undergraduate students of Cardiff University (32 females; 18-28 years of age, M = 20.33, SD = 1.91) participated in this experiment for course credit or monetary compensation.

Materials and procedure. The materials were the same as in Experiments 1 and 2. The procedure was the same as in Experiment 2, with a single exception. The 'drawing' orienting question was substituted with a 'Cardiff' question (presented in the same red font as the 'drawing' question in Experiments 1 and 2): for each word studied with that task, participants had to rate on a 1-4 scale how often they could encounter it in Cardiff (from 'very rarely' to 'very often'). Consequently, both 'drawing' tests were replaced with 'Cardiff' tests (*C-tests*). Foils first encountered on a C-test 1 and later re-used on test 2 are referred to as *C-foils*.

Results

Accuracy. The descriptive statistics for *d'* are presented in Table 1. A 2 (block: first, second) x 2 (test: P-test, C-test) repeated-measures ANOVA on *d'* scores for studied words revealed that the main effect of block was significant, $F(1, 39) = 32.820$, $MSE = .36$, $p < .001$, $\eta_p^2 = .457$, with performance decreasing from the first ($M = 2.23$, $SD = 0.79$) to the second block ($M = 1.68$, $SD = 0.74$). The main effect of test and the interaction were not significant, both Fs < 1.

Foil endorsement rates. The results for foil endorsement rates are presented in the top-right panels of Figure 2 (combined data) and Figure 3 (split by foil and test type). The matched > mismatched difference between foil endorsement rates persisted despite the change of one experimental task, $t(39) = 3.827$, $SE = .01$, $p < .001$, $d = 0.41$. A sign test again confirmed the pervasiveness of the matched > mismatched pattern in the overall data, $z = 2.028$, $p = .043$ (the frequencies are presented in Table 3).

To further investigate the data, we performed a 2 (foil type: P-foil, C-foil) x 2 (test 2 type: P-test, C-test) repeated-measures ANOVA. The main effect of foil was not significant, and neither was the main effect of test 2, $F < 1$ and $F(1, 39) = 1.424$, $MSE = .002$, $p = .24$, $\eta_p^2 = .035$, respectively. There was, however, a significant interaction, $F(1, 39) = 15.002$, $MSE = .006$, $p < .001$, $\eta_p^2 = .278$. C-foils were endorsed as targets more often when presented on a C-test than on a P-test, $t(39) = 2.744$, $SE = .02$, $p = .009$, $d = 0.46$. For P-foils, the analogous difference was not significant, $t(39) = 1.857$, $SE = .02$, $p = .071$, $d = 0.27$, although the numerical trend was in the predicted direction.

Discussion

In this experiment, we again replicated the overall increase in false alarms to non-studied foils subjected twice to monitoring with the same rather than different retrieval orientations. We also demonstrated that this effect is more robust than the results of Experiments 1 and 2 could suggest: it is not specific to the pleasantness task, but can be found in other deep orienting tasks as well. Whereas in Experiments 1 and 2 this pattern of increased false alarm rates for foils repeatedly monitored in the same way was present in the pleasantness task, the present experiment revealed this pattern in a novel Cardiff task.

One curious aspect of the results of Experiment 3 is that this time a reliable matched > mismatched difference failed to materialize for the pleasantness task. Given the results of Experiments 1 and 2, it is clear that this failure cannot be due to the nature of the pleasantness task itself. It does remain surprising, however, that the full predicted pattern of costs of matching monitoring across different tasks was absent from all our experiments.

Assuming that this apparent lack of consistency of monitoring costs is not due to a sampling error, there are at least two related potential explanations. The first explanation posits that monitoring processes at retrieval may not be flexible enough to accommodate two different retrieval orientations within a single memory task. All previous investigations of retrieval orientation with the memory-for-foils paradigm focused on a comparison of deep and shallow processing tasks. Importantly, Danckert et al. (2011) showed that retrieval orientation is likely to be adopted only for a deep test, but not necessarily on a shallow test. As mentioned earlier, in their study, foils from the deep and shallow

tests were subjected to the same judgments that defined deep and shallow tests. Whereas facilitation in the form of quicker responding was found for deep foils, it was absent for shallow foils. These results indicate that all previous studies using memory-for-foils paradigm effectively asked participants to use a single retrieval orientation. By contrast, as our paradigm used two deep tasks, we expected it to create conditions in which participants would adopt two different retrieval orientations. If, however, monitoring processes are not flexible enough to accommodate two different retrieval orientations, it is viable that the costs of monitoring would emerge for only one, presumably more distinctive task.

Another possibility is that the use of only one retrieval orientation does not result from the inflexibility of monitoring processes but instead reflects participants' strategic choices. The use of two different retrieval orientations in our study could have been obviated by the availability of alternative monitoring strategies. It is widely assumed that participants can choose to monitor their output on memory tests by using late corrections for already retrieved memories (see e.g., Halamish et al., 2012). One particular example of late correction is a recall-to-reject approach (e.g., Rotello & Heit, 1999). By this account, on a deep test participants try to retrieve all available details for all test items, including details afforded by encoding with the other, non-target deep task, and then use the details diagnostic of the other task to reject other-task foils. It is possible that in our paradigm participants would use retrieval orientation on the test for the presumably more distinctive task and then strategically switch to a recall-to-reject strategy when other-task foils afford distinctive recollections.

The possibility that recall-to-reject contributed to our results poses also problems for the interpretation of our main finding of an overall increase in false alarms to matched, as compared to mismatched, foils. Thus far we interpreted this finding, obtained for only one retrieval orientation in Experiments 1-3, as indicating that memory monitoring by retrieval orientation can render foils more similar to targets and thus more confusable with these targets when subjected to monitoring for the second time with the same retrieval orientation. We argue that this confusion increases false alarms for matched foils, revealing how memory monitoring can backfire. However, since this argument rests solely on a comparison of matched and mismatched conditions, one could argue that the same pattern would obtain if enriched foils were more effectively rejected in the mismatched condition. The use of the recall-to-reject strategy at test could lead to exactly such a pattern of results. If foils become enriched with details associated with one retrieval orientation and thus confused with targets encoded in the corresponding orienting task, then these new foils, when repeated on the second test for targets studied with the other orienting task, could be more often (erroneously) classified as other-task foils and rejected via the recall-to-reject mechanism. To refute such an argument, an experiment is needed in which monitoring by rejecting items based on recollections they elicit (i.e. recall-to-reject) would be rendered ineffective.

Experiment 4 was designed to remove any reason for the use of a recall-to-reject strategy in the paradigm used for Experiments 1-3. This served two purposes. First, doing so could encourage participants to use monitoring by retrieval orientation for two different deep tasks, potentially revealing the

pattern of costs due to repeated monitoring for both tasks, which was missing from Experiments 1-3. However, were we to replicate the asymmetry documented in these previous experiments even when recall-to-reject was not an effective strategy, it could be concluded that monitoring strategies are inherently inflexible and do not allow for monitoring with two different retrieval orientations. Second, discouraging recall-to-reject means that a replication of the overall matched > mismatched pattern could be more confidently assigned to the costs of repeated monitoring of foils with the same retrieval orientation.

In Experiment 4, we followed an approach of disabling monitoring by recall-to-reject taken from previous studies of retrieval monitoring strategies (e.g., Gallo, Weiss, & Schacter, 2004; McDonough & Gallo, 2012). A novel feature of the procedure, introduced to render recall-to-reject unfeasible as a successful monitoring strategy, was an inclusion of an additional set of words at study which were encountered in both encoding tasks. These repeated words should become associated with details diagnostic of study with both encoding tasks. Crucially, the instructions for the exclusion tasks were also amended to explicitly refer to the presence of repeated words in the exclusion tests. These amended instructions mentioned that since some of the words were studied with both encoding tasks, recollecting that a given word was studied with one task could not be taken as evidence that it was not studied with the other task. As a result, recollections should not feed into the recall-to-reject strategy. For example, participants taking the C-test and encountering a word that elicits recollections of being studied with the P-task should not take these recollections to automatically imply that a given word is an other-task foil. As a result, we would

argue that the most viable monitoring strategy is adopting a retrieval orientation and rejecting items that fail to elicit recollections of being studied with the C-task.

Experiment 4

Method

Participants. Thirty-eight undergraduate students from Cardiff University (37 females; 17-22 years of age, M = 19.05, SD = 1.14) participated in this experiment for course credit. Due to experimenter error, no data for the first C-test (in block 1) were recorded for six participants.[4]

Materials and procedure. The same materials were used as in Experiments 1-3. The composition of the study and test lists, however, had to be changed to accommodate the addition of items presented with both orienting tasks. Each study list consisted of 144 unique words: 124 of these words were presented with one orienting question only, as in Experiments 1-3, whereas the remaining 20 were presented with both orienting questions. This increased the number of study item presentations to 164. Within each test list, there were 62 words studied with a single orienting task (of which 31 served as targets and 31 as to-be-rejected foils from the other orienting task), 10 words studied with both tasks, and 72 foils.

The procedure was the same as in Experiments 1-3, with the following exceptions. Before the study phase, participants were informed that some words

[4] This affected only *d'* analyses which compared block-1 and block-2 *d'* scores; analyses of foil endorsement were performed on block-2 data only, and therefore remained unaffected.

would be presented with both orienting questions, and test instructions were modified to indicate that a word studied in both tasks was always to be endorsed as target.

Results

Accuracy. The descriptive statistics for *d'* are presented in Table 1. A 2 (block: first, second) x 2 (test: P-test, C-test) repeated-measures ANOVA performed on the data of 32 participants who provided results for all tests revealed only a significant main effect of block, $F(1, 31) = 23.870$, $MSE = .152$, $p < .001$, $\eta_p^2 = .435$, with accuracy being higher in the first block ($M = 1.69$, $SD = 0.78$) than in the second block ($M = 1.35$, $SD = 0.82$). The main effect of test and the interaction were not significant, both Fs < 1.

Foil endorsement rates. The bottom-right panels of Figure 2 and Figure 3 present the data for foil endorsement rates. The matched > mismatched difference was again significant, $t(37) = 3.926$, $SE = .009$, $p < .001$, $d = 0.64$, and consistent across participants as assessed by a sign test, $z = 3.381$, $p = .001$ (see Table 3 for the frequencies). When the data were split by task type, a 2 (foil type: P-foil, C-foil) x 2 (test type: P-test, C-test) repeated-measures ANOVA revealed a significant main effect of test, $F(1, 37) = 7.737$, $MSE = .004$, $p = .008$, $\eta_p^2 = .173$, which was qualified by a significant test x foil interaction, $F(1, 37) = 16.253$, $MSE = .003$, $p < .001$, $\eta_p^2 = .305$. The interaction arose because C-foils were more often mistakenly endorsed as targets on a C-test than on a P-test, $t(37) = 4.744$, $SE = .013$, $p < .001$, $d = 0.77$. For P-foils, however, the same pattern was not reliable, $t < 1$.

Discussion

The results of the present experiment replicate the results of Experiments 1-3. Once again, overall false alarm rates were higher for matched foils, presented twice in the same type of test, than for mismatched foils, which were presented in two different types of test. Importantly, this difference again emerged only for foils taken from one type of the exclusion task – the Cardiff task, but not the pleasantness task, as in Experiment 3.

The novel feature of the present experiment was the inclusion of items that were encoded with both deep orienting tasks. Such items have been previously used in studies on monitoring processes to discourage monitoring via a late correction process of recall-to-reject (e.g., Gallo et al., 2004). Given that participants in the present experiment could not have concluded from recollecting other-task details that a given recognition item is an other-task foil, they should have refrained from using recall-to-reject at test. The fact that an asymmetry in monitoring costs across our two orienting tasks was again observed suggests that inconsistent use of retrieval orientations was not due to participants' strategic choice of relying on the recall-to-reject strategy but rather due to inflexibility of monitoring processes via retrieval orientations. Furthermore, the fact that the overall matched > mismatched pattern was again obtained in the present experiment implies that the same patterns observed in Experiments 1-3 were unlikely to be caused by more efficient recall-to-reject processes for new foils repeated across two different types of exclusion tasks. Instead, this pattern, which occurs even when recall-to-reject is rendered

useless, reflects less efficient monitoring via retrieval orientation of foils repeated within the same type of an exclusion task.

Two issues require additional discussion in relation to our present results. The first one concerns the aforementioned asymmetry between the two orienting tasks. Although the present results are consistent with the results of Experiment 3, showing the matched > mismatched pattern for C-foils and not for P-foils, one should remain careful when interpreting null results. In such circumstances, Bayesian tests can be used to assess the strength of evidence for either the experimental or null hypothesis, with the value of BF = 3 considered by some to be a good cut-off for 'substantial' amount of evidence (e.g., Wetzels & Wagenmakers, 2012). When a paired-samples Bayesian t-test was performed on P-foil results of Experiment 3, using the JASP software (JASP Team, 2016), the resulting Bayes factor provided only anecdotal evidence for the null hypothesis, $BF_{01} = 1.45$. However, the same analysis performed for the present experiment returned $BF_{01} = 5.09$, pointing to substantial evidence supporting the null hypothesis.[5] These results indicate that at least under conditions of the present study, designed to remove influences of the recall-to-reject strategy, the matched > mismatched pattern is observed reliably for a single foil type only (C-foils, for which the corresponding Bayes factor in favor of the experimental hypothesis was 731.80).

The other issue concerns a possible caveat of the present procedure, that is the number of targets studied with both orienting tasks. In our design

[5] It is worth noting here that the same Bayesian t-test performed on P-foil results from Experiments 1 and 2 returned $BF_{10} = 17.32$, which can be interpreted as strong evidence in favor of the experimental hypothesis.

participants were presented with only 20 repeated items out of the total of 144 items presented in a given study list, a ratio lower than in previous studies using this technique of disabling recall-to-reject (e.g., Gallo et al., 2004; McDonough & Gallo, 2012). In theory, it could be argued that our method for disabling recall-to-reject was therefore ineffective and participants in Experiment 4 still viewed the recall-to-reject strategy as viable, as it would lead to incorrect inferences for only a small minority of test trials.[6] However, a look at raw false alarm rates for other-task foils, presented in Table 2, suggests that the addition of items studied in both tasks clearly led to a strategy change. The mean false alarm rate for these foils (collapsed across tasks and blocks) for Experiments 1-3 was .14, which rose to .26 in Experiment 4.[7]

This increase in false alarm rates for other-task foils could be explained by the fact that recall-to-reject was no longer a useful strategy in Experiment 4. In Experiments 1-3, even though participants by and large were monitoring by retrieval orientation (as stems unambiguously from the pattern of results), on a small subset of trials the recollection of studying the word with the other task might have been more readily available. As there was no overlap between the study lists presented with the two orienting tasks, these words could then be easily rejected as foils without the need for being monitored for the presence of target-like qualities, lowering false alarms as a result. In Experiment 4, on the other hand, the information about the item being studied in the other task was

[6] We thank an anonymous reviewer for raising this issue.
[7] A formal analysis of false alarm rates to other-task foils for Experiments 3 and 4 (which used the same tasks) with a 2 (Experiment) x 2 (Block) mixed ANOVA yielded only a significant effect of Experiment, $F(1, 76) = 17.32$, $MSE = .025$, $p < .001$, $\eta_p^2 = .19$, which confirms a robust increase in false alarm rates for these foils in Experiment 4.

no longer diagnostic, as the test list contained items studied in both tasks, and monitoring by retrieval orientation was a useful strategy on all trials.

If it is assumed that recall-to-reject was used in Experiments 1-3, but not in Experiment 4, then a question remains why it affected only false alarm rates to other-task foils, while rejection rates for mismatched foils in Experiments 3 and 4, which used the same orienting tasks, were virtually identical (for P-foils, .11 in Experiment 3 and .14 in Experiment 4, and for C-foils, .10 in both experiments). The simplest explanation is that, in order for recall-to-reject to override the dominant strategy of monitoring by retrieval orientation, the memory for studying the word with the other orienting task needs to be easily accessible. This might be the case for some of the other-task foils, but less likely for mismatched foils, which were never intentionally studied with the non-target task. This should not be taken to imply that recall-to-reject could not have been used for mismatched foils in Experiments 1-3, but rather that its contribution likely would have been more limited.

Overall, our manipulation of including items studied with both tasks in study and test lists is likely to have been successful in restricting the operation of the recall-to-reject strategy in Experiment 4 without changing the overall pattern of results. This strongly suggests that recall-to-reject is not necessary for the matched > mismatched pattern to be revealed in the data.

General Discussion

In four experiments we assessed the consequences of repeated monitoring of the same foils with the same retrieval orientation. We found that,

on average, non-studied foils were more often mistaken for targets when there was an overlap between the kinds of information required for target endorsement on tests on which the foil was encountered for the first and second time, than when the tests required retrieval of different kind of information. We interpret this finding as reflecting costs associated with monitoring by retrieval orientation on tests querying for deeply encoded information. By this account, when a non-studied foil is assessed on the first deep test, it is processed in the context of the orienting task in a way similar (even though not identical) to that in which targets were processed at study, and starts resembling items actually studied with that orienting task (e.g., Danckert et al., 2011; Herron & Rugg, 2003a; Jacoby, Shimizu, Velanova, & Rhodes, 2005). When, subsequently, monitoring of that foil is again required on the same type of test, the foil becomes more difficult to reject than equally familiar foils first encountered on the other type of test, as now it shares some characteristics with the searched-for targets – a feature that foils from the other test do not possess.

Are there other plausible explanations for our findings? One important aspect of our procedure is that on each test trial, to-be-assessed words were accompanied by a cue reminding participants which test they were completing ('drawing', 'pleasantness', or 'Cardiff'). These cues were presented in the same place and color as their respective orienting questions at study. It is therefore possible that, when presented with a non-studied foil on the first test, participants associated the foil with the test cue. If such a foil-to-cue association was encoded, this could affect responding on the second test: if the same cue was used on both tests, as in the case of matched foils, on test 2 it would in effect

reinstate the context in which the foil was first encountered. As context reinstatement can boost memory in recognition tests even when participants are not asked to associate items with contexts at study (e.g., Hanczakowski, Zawadzka, & Coote, 2014; Hanczakowski, Zawadzka, & Macken, 2015), reinstating context could have led to higher endorsement rates for matched foils than merely presenting mismatched foils in a familiar context (i.e. with a cue from the other orienting task). At a first pass, this context explanation can account for the matched > mismatched pattern found in the overall data set. However, it cannot explain the fact that the matched > mismatched difference in endorsement rates was noticeably more pronounced for foils first encountered on one type of test (P-test in Experiments 1 and 2, C-test in Experiments 3 and 4) than the other. There is no *a priori* explanation for why participants would benefit from reinstatement of one incidental context, but not the other. We therefore conclude that context reinstatement is not a viable explanation of our results.

Another possibility worth considering is whether our results can be explained by test-dependent differences in criterion setting. If, for example, participants' responding in Experiments 1 and 2 was more liberal on P-tests than on D-tests, overall more foils (both matched and mismatched) would be endorsed on P-tests. This would lead to a greater difference between foil endorsement rates *between* P- and D-tests for P-foils than for D-foils - a pattern exactly like the one found in our data. It has to be noted, however, that if such differences in criterion setting were indeed occurring, they would be revealed as the main effect of test type in 2 (foil type) x 2 (test type) ANOVAs for foil

endorsement rates which were reported for each of the experiments. As this main effect was significant only in Experiment 4, but not in Experiments 1-3, it is unlikely that differences in criterion setting can account for our results.

Finally, if assessing non-studied foils at test 1 indeed enriches them with characteristics resembling those possessed by targets, as predicted by the retrieval orientation account, there are two fundamentally different factors that can build upon this to produce the matched > mismatched pattern we consistently found on test 2 in all experiments. The first option is that matched foils are more often mistakenly endorsed as targets than mismatched foils, as participants detect the features that matched foils share with targets. This explanation assumes that on test 2 the retrieval orientation that guides participants' old/new decisions leads to impaired monitoring of matched foils. According to the second option, mismatched foils are more often correctly rejected than matched foils, as participants are able to recollect information that (incorrectly) suggests that those items were studied with the other orienting task. By virtue of this recall-to-reject process (e.g., Rotello & Heit, 1999), participants would then be more able to improve their monitoring of repeated foils on the second test.

The recall-to-reject account of our results might be seen to be challenged by the fact that ERP studies have shown that when memory for targets is strong, the left parietal old/new effect - commonly interpreted as an index of recollection (see Rugg & Allan, 2000, for a review) - for foils studied with the other orienting task is either absent (Dzulkifli & Wilding, 2005; Herron & Rugg, 2003b) or markedly attenuated compared to the situation when target memory

is weaker (Bridger et al., 2009). This suggests that the role of recollection of other-task information in exclusion tests could be limited, at least when targets are strongly encoded. The strategy of actively searching for target-specific information, on the other hand, has consistently found its reflection in the results of studies suggesting operation of different retrieval orientations depending on the orienting task - in ERP data (e.g., Bridger et al., 2009; Dzulkifli & Wilding, 2005; Herron & Rugg, 2003a), behavioral data (e.g., Danckert et al., 2011; Halamish et al., 2012), and self-report (Herron & Rugg, 2003b). Note, however, that if it was assumed that recall-to-reject operated in Experiments 1-3 (although not in Experiment 4), this suggests that those ERP findings are not generalizable to the paradigm used in our study.

The possibility that a recall-to-reject strategy was adopted by participants on some of the block-2 test trials in Experiments 1-3 motivated our direct attempt to eliminate recall-to-reject processing in Experiment 4, which left the crucial results of this study unchanged. In this experiment items studied with both orienting tasks were included and participants were told that recollecting other-task details is not informative with regard to target/foil status of a recognition item (see Gallo et al., 2004, for the same methodology). Under these conditions, the recall-to-reject strategy is clearly a less optimal monitoring strategy than monitoring by retrieval orientation. The fact that results of Experiment 4 remain consistent with the results of Experiment 1-3 suggests that the discussed matched > mismatched pattern should be assigned to the after-effects of monitoring by retrieval orientation.

One surprising aspect of our results is that although all four experiments revealed the matched > mismatched pattern in overall false alarm rates, a more detailed analysis of false alarms at the level of individual tasks shows that across experiments the matched > mismatched pattern was always observed only in one of the tasks. It is crucial to stress that this asymmetry cannot be due to the particular tasks we used in our study. Whereas Experiments 1 and 2 revealed the matched > mismatched pattern for the pleasantness but not for the drawing task, Experiments 3 and 4 revealed it for the Cardiff but not for the pleasantness task. It seems plausible that what matters for this particular asymmetry is the relative rather than absolute distinctiveness of the particular tasks used in the experimental procedure.

The asymmetry observed in our results is best accounted for by assuming that participants adopt only a single retrieval orientation in a given experimental task.[8] This willingness to adopt only a single orientation seems to be unrelated to the availability of an alternative monitoring strategy in the form of recall-to-reject: in Experiment 4, the role of recall-to-reject was minimized, yet the asymmetry was still clearly observed. Thus, it seems that this asymmetry reflects more general limits on flexibility of adopting or adjusting retrieval orientation. The problem of flexibility of retrieval orientations was previously examined by Marsh et al. (2009) who, using the basic memory-for-foils paradigm (see Introduction), examined participants' ability to switch retrieval orientation on a trial-by-trial basis. Although the usual pattern of memory-for-foils was observed

[8] Note that our measure of false alarms depends on retrieval orientations for both the first and second blocks of testing. Our paradigm does not allow for assessing whether a single retrieval orientation was adopted in the first block, second block or in both blocks.

in this study, which was interpreted as evidence of flexibility in the use of retrieval orientations, the particular paradigm chosen by Marsh et al. required a comparison between deep and shallow orienting tasks. As argued by Danckert et al. (2011), one can legitimately describe a deep retrieval orientation but not a shallow retrieval orientation, which suggests that perhaps what Marsh et al. examined was not so much switching between retrieval orientations but switching on and off a single retrieval orientation.

To our knowledge, the only experiment to employ the memory-for-foils paradigm and two deep orienting tasks was conducted by Alban and Kelley (2012, Experiment 2). In this experiment, participants studied lists of words with two shallow tasks (counting curved letters and counting vowels) and two deep tasks (a pleasantness task and a task of assessing whether an item would fit into a shoebox). The memory-for-foils paradigm was employed for words studied with one shallow (counting vowels) and one deep task (the shoebox task) but, crucially, the relevant tests were preceded by either a recognition task for words from the other shallow or from the other deep task. Participants for whom the main task was preceded with a test for shallowly encoded items showed the usual memory-for-foils pattern but participants for whom the main task was preceded with a test for deeply encoded items did not. Of main interest is the latter group which was effectively tested for words studied with two different deep tasks and who failed to show evidence for the memory-for-foils effect.

This pattern was interpreted by Alban and Kelley (2012) as indicating that an easy recognition task for deeply encoded items prevents participants from adopting a retrieval orientation in the main task. However, an alternative

possibility is that a retrieval orientation was in fact adopted on the pleasantness test and not for the later shoebox task, consistent with our hypothesis of inflexibility of retrieval orientations. This possibility cannot be assessed because Alban and Kelley tested foils only from the main task and not from the pleasantness test that preceded it. The central point remains, however, which is that the only behavioral study conducted to date that included tests for two deep orienting tasks failed to find evidence for the adoption of retrieval orientation for the only task for which this issue was assessed. In conclusion, there is currently a conspicuous lack of behavioral evidence for participants using two different retrieval orientations within a single memory task. This constitutes a clear direction for further studies that could adopt a variety of dependent measures developed within the memory-for-foils methodology – hits for re-presented foils (e.g., Jacoby, Shimizu, Velanova, & Rhodes, 2005), speed of answering the orienting question for previous foils (Danckert et al., 2011), and false alarms for matching/mismatching foils (as in the present study) – to investigate the flexibility of early-selection monitoring processes.

On a broader note, the present results relate to a wide topic of repeated processing of the same items within a memory task. Repeated presentations of the same item within a test has been shown to have a detrimental effect on accuracy of memory judgments by increasing false alarms to repeated foils (Jennings and Jacoby, 1997). The present study suggests that such negative effects of foil repetition can be at least to some extent remedied if foils are processed differently on each occasion on which they are encountered. These benefits of processing variability are reminiscent of similar ideas revealed in

studies of encoding. Recently, Huff and Bodner (2014) re-examined the old issue of encoding variability and showed that what ultimately matters for effective encoding is the variability in processing to which repeatedly presented study items are subjected. The apparent similarities between the roles of encoding and retrieval variability in supporting accurate remembering are perhaps unsurprising given the latest focus on how encoding and retrieval processes are interleaved and thus at least to some extent governed by similar principles (see, e.g., Tullis, Benjamin, & Ross, 2014).

In the present study we have demonstrated adopting a retrieval orientation on an exclusion test - a strategy which normally is beneficial to test performance (although see Kantner & Lindsay, 2013) - can lead to impaired monitoring when repeated assessment of the same foils is required. Potential costs of monitoring by retrieval orientation have previously been suggested by Gray and Gallo (2015), and in the present study we have shown when and how these costs can emerge. These results join other findings demonstrating negative aspects of employing generally beneficial strategies of learning and testing. For example, although retrieval practice in most cases leads to memory improvement as compared to restudying the to-be-learned material (e.g., Roediger & Karpicke, 2006), it can also hinder performance when it diverts attention from the structure of the study list (Peterson & Mulligan, 2013). Also, restudying the same stimuli under certain conditions can impair performance as compared to a single study episode (Mulligan & Peterson, 2013, 2014), and orthographic distinctiveness, known to produce benefits in recall (e.g., Geraci & Rajaram, 2002), can also lead to an impairment in encoding inter-item

associations at a cost to memory performance (McDaniel, Cahill, & Bugg, 2015). We believe that investigating those exceptions to general patterns is necessary both from the theoretical and practical point of view. For theory, exceptions can be crucial for determining the mechanisms behind common findings. Also from the perspective of learners, it is important not only to know which strategies can be used to boost memory performance, but also when it is best to avoid using them. The present results further our understanding of foil processing at test and offer an important step toward understanding its practical consequences.

References

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

Benjamin, A. S., & Bawa, S. (2004). Distractor plausibility and criterion placement in recognition. *Journal of Memory and Language, 51,* 159-172. http://dx.doi.org/10.1016/j.jml.2004.04.001

Bridger, E. K., Herron, J. E., Elward, R. L., & Wilding, E. L. (2009). Neural correlates of individual differences in strategic retrieval processing. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35*, 1175–1186. http://dx.doi.org/10.1037/a0016375

Coltheart, M. (1981). The MRC Psycholinguistic Database. *Quarterly Journal of Experimental Psychology Section A, 33,* 497-505. http://dx.doi.org/10.1080/14640748108400805

Danckert, S. L., MacLeod, C. M., & Fernandes, M. A. (2011). Source-constrained retrieval influences the encoding of new information. *Memory & Cognition, 39,* 1374-1386. http://dx.doi.org/10.3758/s13421-011-0117-9

Dodson, C. S., & Schacter, D. L. (2002). When false recognition meets metacognition: The distinctiveness heuristic. *Journal of Memory and Language, 46*, 782-803. http://dx.doi.org/10.1006/jmla.2001.2822

Dzulkifli, M. A., & Wilding, E. L. (2005). Electrophysiological indices of strategic retrieval processing. *Neuropsychologia, 43,* 1152-1162. http://dx.doi.org/10.1016/j.neuropsychologia.2004.11.019

Gallo, D. A., Weiss, J. A., & Schacter, D. L. (2004). Reducing false recognition with criterial recollection tests: Distinctiveness heuristic versus criterion shifts. *Journal*

of Memory and Language, 51, 473-493. http://dx.doi.org/10.1016/j.jml.2004.06.002

Gardiner, J. M. (1988). Functional aspects of recollective experience. *Memory & Cognition,* 16**,** 309-313. http://dx.doi.org/10.3758/BF03197041

Geraci, L., & Rajaram, S. (2002). The orthographic distinctiveness effect on direct and indirect tests of memory: Delineating the awareness and processing requirements. *Journal of Memory & Language, 47*, 273-291. http://dx.doi.org/10.1016/S0749-596X(02)00008-6

Gray, S. J., & Gallo, D. A. (2015). Disregarding familiarity during recollection attempts: Content-specific recapitulation as a global retrieval orientation strategy. *Journal of Experimental Psychology: Learning, Memory & Cognition, 41,* 134-147. doi: http://dx.doi.org/10.1037/a0038363

Guzel, M. E., & Higham, P. A. (2013). Dissociating early- and late-selection processes in recall: The mixed blessing of categorized study lists. *Memory & Cognition, 41,* 683-697. http://dx.doi.org/10.3758/s13421-012-0292-3

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. http://dx.doi.org/10.1037/a0025053

Hanczakowski, M., & Mazzoni, G. (2011). Both differences in encoding processes and monitoring at retrieval reduce false alarms when distinctive information is studied. *Memory*, 19(3), 280-289. http://dx.doi.org/10.1080/09658211.2011.558514

Hanczakowski, M., Zawadzka, K., & Coote, L. (2014). Context reinstatement in recognition: Memory and beyond. *Journal of Memory and Language, 72,* 85-97. doi: http://dx.doi.org/10.1016/j.jml.2014.01.001

Hanczakowski, M., Zawadzka, K., & Macken, B. (2015). Continued effects of context reinstatement in recognition. *Memory & Cognition, 43,* 788-797. http://dx.doi.org/10.3758/s13421-014-0502-2

Herron, J., & Rugg, M. D. (2003). Retrieval orientation and the control of recollection. *Journal of Cognitive Neuroscience*, 15(6), 843-854. http://dx.doi.org/10.1162/089892903322370762

Herron, J., & Rugg, M. D. (2003). Strategic influences on recollection in the exclusion task: Electrophysiological evidence. *Psychonomic Bulletin & Review*, 10(3), 703-710. http://dx.doi.org/10.3758/BF03196535

Herron, J. & Wilding, E. L. (2004). An electrophysiological dissociation of retrieval mode and retrieval orientation. *Neuroimage*, 22(4), 1554-1562. http://dx.doi.org/10.1016/j.neuroimage.2004.04.011

Herron, J. & Wilding, E. L. (2006). Neural correlates of control processes engaged before and during recovery of information from episodic memory. *NeuroImage, 30*, 634–644. http://dx.doi.org/10.1016/j.neuroimage.2005.10.003

Huff, M. J., & Bodner, G. E. (2014). All varieties of encoding variability are not created equal: Separating variable processing from variable tasks. *Journal of Memory and Language, 73,* 43-58. http://dx.doi.org/10.1016/j.jml.2014.02.004

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

Jacoby, L. L., Kelley, C. M., & McElree, B. D. (1999). The role of cognitive control: Early selection vs. late correction. In S. Chaiken & Y. Trope (Eds.), *Dual-process theories in social psychology* (pp. 383-400). New York: Guilford

Jacoby, L. L., Shimizu, Y., Daniels, K. A. & Rhodes, M. (2005). Modes of cognitive

control in recognition and source memory: Depth of retrieval. *Psychonomic Bulletin and Review, 12*, 852-857. http://dx.doi.org/10.3758/BF03196776

Jacoby, L. L., Shimizu, Y. Velanova, K., & Rhodes, M. (2005). Age differences in depth of retrieval: Memory for foils. *Journal of Memory and Language, 52,* 493-504. http://dx.doi.org/10.1016/j.jml.2005.01.007

JASP Team (2016). JASP (Version 0.7.5 Beta 2)[Computer software]

Jennings, J. M., & Jacoby, L. L. (1997). An opposition procedure for detecting age-related deficits in recollection: Telling effects of repetition. *Psychology and Aging, 12,* 352-361. http://dx.doi.org/10.1037/0882-7974.12.2.352

Kantner, J., & Lindsay, D. S. (2013). Top-down constraint on recognition memory. Memory & Cognition, 41, 465-479. http://dx.doi.org/10.3758/s13421-012-0265-6

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

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. http://dx.doi.org/10.1016/j.jml.2009.06.005

McDaniel, M. A., Cahill, M. J., & Bugg, J. M. (2015). The Curious Case of Orthographic Distinctiveness: Disruption of Categorical Processing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. Advance online publication. http://dx.doi.org/10.1037/xlm0000160

McDonough, I. M., & Gallo, D. A. (2012). Illusory expectations can affect retrieval-monitoring accuracy. *Journal of Experimental Psychology: Learning, Memory, and*

Cognition, 38, 391-404. http://dx.doi.org/10.1037/a0025548

Mulligan, N. W., & Peterson, D. (2013). The negative repetition effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 39,* 1403–1416. http://dx.doi.org/10.1037/a0031789

Mulligan, N. W., & Peterson, D. J. (2014). Analysis of the encoding factors that produce the negative repetition effect. *Journal of Experimental Psychology: Learning, Memory & Cognition*, 40, 765-775 . http://dx.doi.org/10.1037/a0035577

Peterson, D. J., & Mulligan, N. W. (2013). The negative testing effect and multi-factor account. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 39,* 1287–1293. http://dx.doi.org/10.1037/a0031337

Pierce, B. H., & Gallo, D. A. (2011). Encoding modality can affect memory accuracy via retrieval orientation. *Journal of Experimental Psychology: Learning, Memory & Cognition*, *37*, 516-521. http://dx.doi.org/10.1037/a0022217

Robb, W. G., & Rugg, M. D. (2002). Electrophysiological dissociation of retrieval orientation and retrieval effort. *Psychonomic Bulletin and Review, 9*, 583-589. http://dx.doi.org/10.3758/BF03196316

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

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

Rotello, C. M., & Heit, E. (1999). Two-process models of recognition memory: Evidence for recall-to-reject? *Journal of Memory and Language, 40,* 432-453.

http://dx.doi.org/10.1006/jmla.1998.2623

Rugg, M. D., & Allan, K. (2000). Event-related potential studies of memory. In E. Tulving & F. I. M. Craik (Eds.), *The Oxford Handbook of Memory* (pp. 521–537). Oxford, UK: Oxford University Press.

Rugg, M. D., Allan, K., & Birch, C. S. (2000). Electrophysiological evidence for the modulation of retrieval orientation by depth of study processing. *Journal of Cognitive Neuroscience, 12*, 664-678. http://dx.doi.org/10.1162/089892900562291

Rugg, M. D., & Wilding, E. L. (2000). Retrieval processing and episodic memory. *Trends in Cognitive Sciences, 4*, 108-115. http://dx.doi.org/10.1016/S1364-6613(00)01445-5

Shimizu, Y., & Jacoby, L. L. (2005). Similarity-guided depth of retrieval: Constraining at the front end. *Canadian Journal of Experimental Psychology, 59*, 17-21. http://dx.doi.org/10.1037/h0087455

Tullis, J. G., Benjamin, A. S., & Ross, B. H. (2014). The reminding effect : Presentation of associates enhances memory for related words in a list. *Journal of Experimental Psychology: General, 143,* 1526-1540. http://dx.doi.org/10.1037/a0036036

Tulving, E. (1985). Memory and consciousness. *Canadian Psychology, 26,* 1-12. http://dx.doi.org/10.1037/h0080017

Wetzels, R., & Wagenmakers, E.-J. (2012). A default Bayesian hypothesis test for correlations and partial correlations. *Psychonomic Bulletin & Review, 19*, 1057-1064. http://dx.doi.org/10.3758/s13423-012-0295-x

Table 1.

Accuracy (d') scores across tests and test types in Experiments 1 - 4. P, D, and C refer to the 'pleasantness', 'drawing', and 'Cardiff' tests, respectively.

	d' - studied foils						*d'* - all foils					
	Test 1			Test 2			Test 1			Test 2		
Experiment and Test Type	P	D	C	P	D	C	P	D	C	P	D	C
Experiment 1	2.30 (0.95)	2.53 (0.93)	-	2.07 (1.16)	2.30 (1.02)	-	2.64 (0.82)	2.90 (0.85)	-	2.26 (1.09)	2.29 (0.91)	-
Experiment 2	2.43 (0.76)	2.40 (0.96)	-	1.72 (0.80)	1.99 (0.97)	-	2.76 (0.62)	2.78 (0.94)	-	1.79 (0.75)	2.00 (0.77)	-
Experiment 3	2.24 (0.85)	-	2.21 (0.73)	1.68 (0.73)	-	1.69 (0.76)	2.67 (0.72)	-	2.50 (0.74)	1.83 (0.77)	-	1.71 (0.70)
Experiment 4	1.66 (0.71)	-	1.71 (0.85)	1.34 (0.89)	-	1.36 (0.75)	2.33 (0.70)	-	2.37 (0.76)	1.68 (0.73)	-	1.67 (0.86)

Table 2.

Hit rates and false alarm rates as a function of block, test, and item type in Experiments 1-4.

	Experiment			
Block, Test, and Item Type	Experiment 1	Experiment 2	Experiment 3	Experiment 4
Block 1				
Test A				
Target - single task	.85 (.12)	.88 (.08)	.84 (.11)	.83 (.11)
Target - both tasks	-	-	-	.90 (.11)
Foil - other task	.09 (.09)	.14 (.16)	.14 (.12)	.26 (.18)
Foil - novel	.04 (.06)	.07 (.15)	.05 (.11)	.04 (.06)
Test B				
Target - single task	.81 (.12)	.83 (.10)	.80 (.12)	.76 (.14)
Target - both tasks	-	-	-	.80 (.16)
Foil - other task	.12 (.10)	.14 (.10)	.15 (.12)	.27 (.15)
Foil - novel	.04 (.06)	.03 (.04)	.03 (.04)	.04 (.06)
Block 2				
Test A				
Target - single task	.75 (18)	.71 (.14)	.71 (.16)	.69 (.19)
Target - both tasks	-	-	-	.72 (.20)
Foil - other task	.10 (.08)	.12 (.09)	.14 (.08)	.27 (.15)
Test B				
Target - single task	.72 (.19)	.67 (17)	.68 (.16)	.65 (.16)
Target - both tasks	-	-	-	.72 (.21)
Foil - other task	.16 (.15)	.15 (.10)	.18 (.13)	.23 (.13)

Table 3.
Number of participants displaying matched>mismatched, matched=mismatched, and matched<mismatched patterns in Experiments 1-4.

	direction of difference		
	matched > mismatched	matched = mismatched	matched < mismatched
Experiment 1	15	3	6
Experiment 2	14	4	6
Experiment 3	24	5	11
Experiment 4	28	3	7
Total	81	15	30

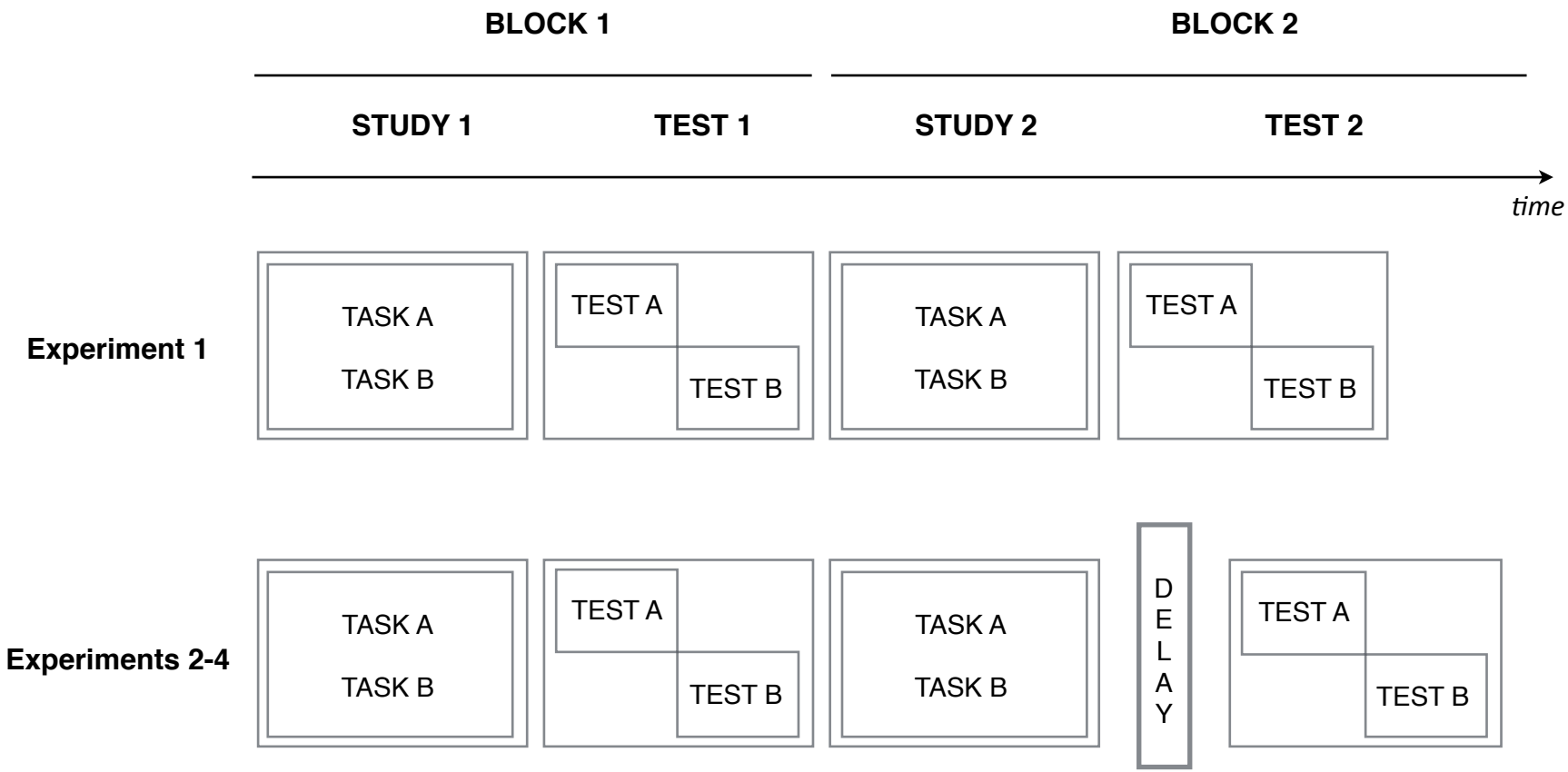

Figure 1. A schematic representation of the design of Experiments 1-4.

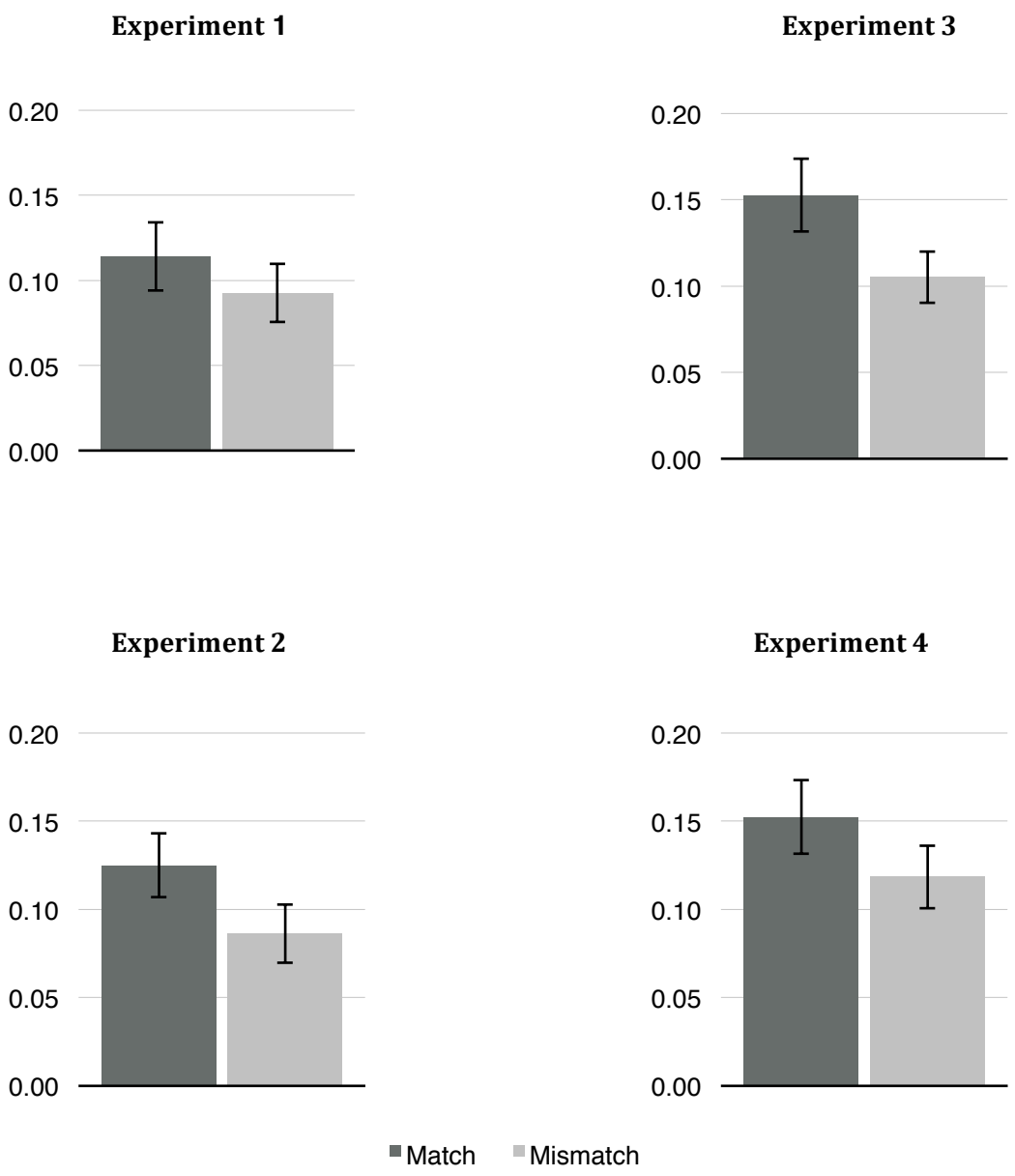

Figure 2. Endorsement rates for non-studied foils as a function of the match between block-1 and block-2 retrieval orientations in Experiments 1 - 4. Error bars denote standard errors.

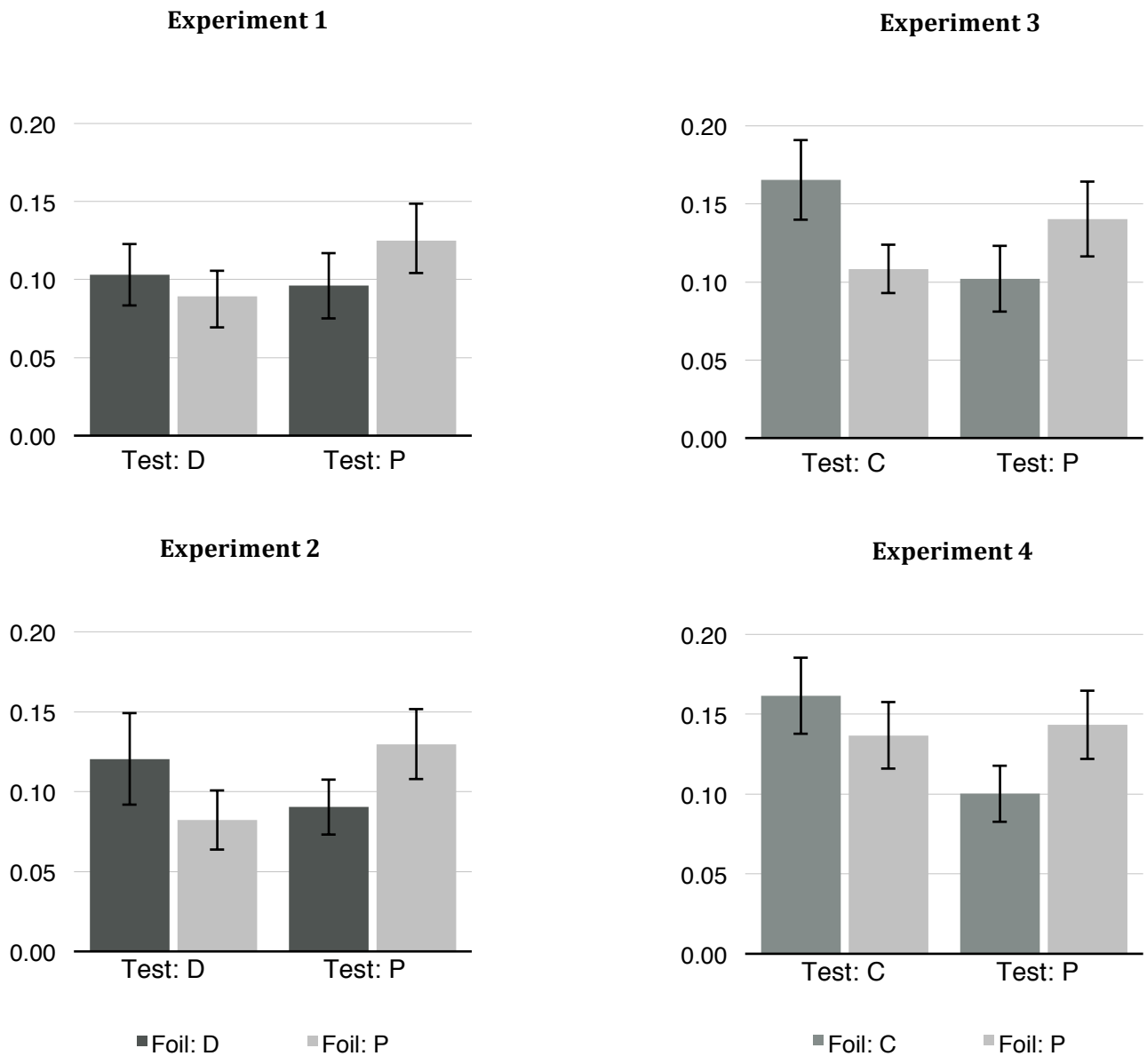

Figure 3. Endorsement rates for non-studied foils as a function of foil and test type in Experiments 1 - 4. P, D, and C refer to 'pleasantness', 'drawing', and 'Cardiff', respectively. Error bars denote standard errors.

Appendix

Analyses of *d'* scores for the full data set including non-studied foils.

Experiment 1

A 2 (block: first, second) x 2 (test: P-test, D-test) repeated-measures ANOVA performed on *d'* scores calculated from the full data set revealed only a significant main effect of block, $F(1, 23) = 29.124$, $MSE = .20$, $p < .001$, $\eta_p^2 = .559$. The main effect of test and the interaction were not significant, $F(1, 23) = 1.665$, $MSE = .32$, $p = .21$, $\eta_p^2 = .068$, and $F(1, 23) = 2.110$, $MSE = .14$, $p = .16$, $\eta_p^2 = .084$, respectively.

Experiment 2

A 2 (block: first, second) x 2 (test: P-test, D-test) repeated-measures ANOVA revealed that the main effect of block was significant, $F(1, 23) = 67.988$, $MSE = .27$, $p < .001$, $\eta_p^2 = .747$, while the main effect of test and the interaction were not, $F(1, 23) = 1.441$, $MSE = .23$, $p = .24$, $\eta_p^2 = .059$, and $F(1, 23) = 1.328$, $MSE = .18$, $p = .26$, $\eta_p^2 = .055$, respectively.

Experiment 3

A 2 (block: first, second) x 2 (test: P-test, C-test) repeated-measures ANOVA revealed that the main effect of block was again significant, $F(1, 39) = 102.570$, $MSE = .26$, $p < .001$, $\eta_p^2 = .725$, while the main effect of test and the interaction were not, both Fs < 1.

Experiment 4

A 2 (block: first, second) x 2 (test: P-test, C-test) repeated-measures ANOVA again revealed only a significant main effect of block, $F(1,31) = 83.625$, $MSE = .174$, $p < .001$, $\eta_p^2 = .730$, Fs < 1 for the main effect of test and the interaction.