Journal of Experimental Psychology:
Learning, Memory, and Cognition
2016, Vol. 42, No. 7, 1090–1111

0278-7393/16/$12.00 http://dx.doi.org/10.1037/xlm0000214

The d-Prime Directive: Assessing Costs and Benefits in Recognition by Dissociating Mixed-List False Alarm Rates

Noah D. Forrin, Brianna Groot, and Colin M. MacLeod
University of Waterloo

It can be difficult to judge the effectiveness of encoding techniques in a within-subject design. Consider the *production effect*—the finding that words read aloud are better remembered than words read silently. In the absence of a baseline, a within-subject production effect in a mixed study list could reflect a benefit of reading aloud, a cost of reading silently, or both. To help interpret within-subject data, memory researchers have compared within-subject and between-subjects designs, with the between-subjects (i.e., pure list) conditions serving as baselines against which the within-subject (i.e., mixed-list) conditions are compared. In the present article, the authors highlight a shortcoming of using this comparison to assess costs and benefits in recognition. Unlike between-subjects experiments where separate false alarm rates are obtained for each condition, the typical within-subject experiment yields a collapsed false alarm rate, which, the authors argue, can potentially bias calculations of memory discrimination (d'). Across 3 experiments that used production as the encoding manipulation, they used a typical mixed-list versus pure-list design (Experiment 1) and then made modifications to this design (Experiments 2 and 3) that yielded separate mixed-list false alarm rates. The results of the latter 2 experiments demonstrated that words that are read aloud in a mixed list have an overall memorial benefit over words that are read aloud in a pure list—both in terms of increased hits *and* reduced false alarms. The authors frame these results in terms of the distinctiveness heuristic.

Keywords: production effect, recognition memory, mixed list, pure list, costs and benefits

There is a memorial advantage to reading aloud relative to reading silently. Building on initial research by Hopkins and Edwards (1972; see also Conway & Gathercole, 1987; Gathercole & Conway, 1988), MacLeod, Gopie, Hourihan, Neary, and Ozubko (2010) found that this phenomenon—which they named the *production effect*—was robust in within-subject designs (in which aloud and silent words are randomly intermixed at study). MacLeod and colleagues acknowledged, however, that the production effect does not necessarily represent a memorial benefit of reading aloud in a mixed-list: It could also reflect a cost imposed on the silently read words, or it could consist of both cost and benefit. Moreover, this situation is not unique to the production effect: It is relevant whenever designs are compared to ascertain benefits versus costs.

Costs and benefits can be difficult to assess in a within-subject experiment because it can be argued that either condition represents the "baseline" against which performance in the other condition is compared (see Jonides & Mack, 1984). To elucidate within-subject effects in memory, researchers have compared within-subject to between-subjects experimental designs (Begg & Roe, 1988; Begg & Snider, 1987; Slamecka & Katsaiti, 1987; Bodner, Taikh, & Fawcett, 2014).[1] When doing so, between-subjects (i.e., pure-list) conditions serve as empirical baselines against which within-subject (i.e., mixed-list) memory performance is compared.

Notably, the results of a within versus between design comparison can provide useful information for understanding the mechanism underlying a given memory phenomenon. A within-subject benefit suggests that a phenomenon may be driven by distinctive processing at the time of encoding (for an overview of distinctiveness theory, see Hunt, 2006, 2013; Hunt & Worthen, 2006). In Hunt's (2006, p. 12) words, *distinctiveness* is the "processing of difference in the context of similarity." Along these lines, the production effect has been argued to be a distinctiveness effect (see Conway & Gathercole, 1987; MacLeod et al., 2010).[2] The process of reading aloud stands out in a mixed study list in which all words (whether read aloud or silently) share common lexical processing. Moreover, individuals can strategically retrieve this distinct record of speech during a recognition task to determine

This article was published Online First January 28, 2016.

Noah D. Forrin, Brianna Groot, and Colin M. MacLeod, Department of Psychology, University of Waterloo.

This research was supported by Natural Sciences and Engineering Research Council of Canada Discovery Grant A7459. We thank Roshan Cherian, Joanna Collaton, Emily Cyr, Sukhdip Grewal, Sabina Kavecka, Madalina Oancea, Deanna Priori, Kristina Schrage, James Siklos-Whillans, and Madison Stange for their assistance in collecting the data.

Correspondence concerning this article should be addressed to Noah D. Forrin, Department of Psychology, University of Waterloo, 200 University Avenue West, Waterloo, Ontario N2L 3G1, Canada. E-mail: nforrin@uwaterloo.ca

[1] Meta-analytic research has also compared the results of within-subject effects and their between-subjects counterparts (e.g., Bertsch, Pesta, Wiscott, & McDaniel, 2007; Fawcett, 2013; McDaniel & Bugg, 2008).

[2] Other mnemonics have also been cast as distinctiveness effects, including enactment (Engelkamp & Zimmer, 1997), generation (Begg, Snider, Foley, & Goddard, 1989), and bizarreness (McDaniel & Einstein, 1986).

whether a given word is "old" or "new" (see Dodson & Schacter, 2001).

Costs, on the other hand, suggest that an apparently beneficial situation may be illusory, arising from participants' lazy processing of items from the seemingly less important condition. Indeed, Begg and Snider (1987) concluded that the generation effect was an experimental artifact on the basis of a within versus between experiment that revealed a cost for reading aloud (the comparison condition) in a mixed list, but no corresponding benefit of generation, and no between-subjects generation effect (see also Begg & Roe, 1988; Slamecka & Katsaiti, 1987). However, in a subsequent within versus between experiment in which participants generated or read word pairs, Begg, Snider, Foley, and Goddard (1989) found the opposite trend—a benefit of generation without a cost to reading aloud—which led them to proclaim that the generation effect "is no artifact." Generated words, the authors argued, benefited from distinctive processing in a mixed list.

As this generation effect research illustrates, within-subject versus between-subjects designs have been used to better understand the mechanisms underlying memory phenomena—and even to test their legitimacy. We wish to pursue the core issues of these designs, and have chosen to use the production effect as the testing ground. With that as our goal, we will first set the context in this introduction by providing an overview of production effect research—in particular, recent work by Bodner et al. (2014)—that has used within-subject versus between-subjects comparisons to examine the costs and benefits of production. Does production in fact represent a benefit to reading aloud, or merely a cost to reading silently? After reviewing this research, we will argue that there are inherent problems with using the standard within-subject versus between-subjects design comparison to gauge costs and benefits in recognition tasks.

Assessing Costs and Benefits in Production

In the seminal production experiment, Hopkins and Edwards (1972) collected both within-subject and between-subjects recognition data. When they compared their results across these designs, they found a within-subject effect but no between-subjects effect of production. Importantly, Hopkins and Edwards further observed no advantage to reading aloud in a within-subject design relative to reading aloud *or* to reading silently in a between-subjects design. Thus, relative to these between-subjects baselines, there was no benefit of reading aloud. The authors concluded that "the effect of pronunciation appears to lie primarily in a decrement in performance for unpronounced words rather than an increment for recognition memory of pronounced words" (p. 537).

Although the Hopkins and Edwards (1972) results initially suggested that the production effect may simply reflect a cost of reading silently in a mixed list, recent research has cast a more favorable light on the production effect. In particular, meta-analytic work by Fawcett (2013) revealed a significant between-subjects production effect, an effect that had eluded detection in earlier production research that examined hit rates in recognition memory (e.g., MacLeod et al., 2010). Fawcett's meta-analysis, however, used a signal detection measure of memory discrimination (d'), which took into account both hits and false alarms (FAs). The combined benefit of increased hits and decreased FAs in the pure-aloud group resulted in significantly better performance than in the pure-silent group—a between-subjects production effect.

In a recent article, Bodner et al. (2014) used a within-subject versus between-subjects design comparison to assess the costs and benefits of the within-subject production effect in recognition memory. Their research built upon the earlier work of Hopkins and Edwards (1972) by measuring memory discrimination (d'), not simply hits. Bodner et al. also included a meta-analysis that compared all available within-subject versus between-subjects production experiments. Their experiment and their meta-analysis revealed a consistent pattern of data—significant within-subject and between-subjects production effects in memory discrimination (d'). They also found significantly greater discrimination in their mixed-aloud group than in their pure-silent group (which they termed *benefits-over-silent*). These results demonstrated that the production effect is not merely an artifact: Reading aloud does, in fact, enhance memory for words, irrespective of whether one reads aloud in a mixed list or a pure list.

Importantly, however, Bodner et al. (2014) did not find a memorial benefit to reading aloud in a mixed list versus a pure list. Rather, they found a significant mixed-list cost, with recognition of silent words poorer in a mixed list relative to a pure list. This cost was eliminated when the authors used a blocked design, a finding that they noted (p. 5) was consistent with participants "lazily" reading silent words in a mixed study list (see Begg & Snider, 1987). Overall, the pattern of results obtained by Bodner and colleagues—in particular the lack of a mixed-list benefit of reading aloud—runs counter to the claim that production results from a distinctiveness effect (Conway & Gathercole, 1987; MacLeod et al., 2010).

Although we endorse Bodner and colleagues' (2014) approach of comparing within-subject and between-subjects designs to assess the costs and benefits of within-subject production, we submit that, as ordinarily computed, d' is a problematic measure for making comparisons across these experimental designs. The basis for our claim is that FAs in a within-subject experiment have a different meaning than they do in a between-subjects experiment. In a between-subjects production experiment, separate pure-aloud and pure-silent FA rates are obtained from participants who studied pure aloud versus pure silent lists. These separate FA rates permit the calculation of separate pure-aloud and pure-silent d' values. But in a standard within-subject design, it is not possible to obtain separate mixed-aloud and mixed-silent FA rates. There is a single mixed-list FA rate because there is no way of determining whether a given FA occurred because a new word was mistaken as an aloud word or as a silent word.

Essentially, then, the lack of separate FAs for the aloud and silent conditions of a mixed-list design precludes the unbiased calculation of independent mixed-aloud and mixed-silent d' values. In computing aloud and silent d' values in the mixed design, Bodner et al. (2014) had to use the single overall mixed FA rate as an estimate both for the proportion of new words misclassified as aloud (FA aloud) and for the proportion of new words misclassified as silent (FA silent). Consequently, the two d' rates were not independent.

We contend that this approach is problematic because it presumes that individuals, if asked to make modality attributions on a recognition test (i.e., to classify test items as *aloud*, *silent*, or *new*), would be as likely to misclassify new words as aloud as they

would be to misclassify new words as silent. Contrary to this assumption, within-subject production experiments that have asked participants to make modality attributions at test have consistently demonstrated that participants are significantly less likely to miscategorize a new word as aloud than as silent (see Conway & Gathercole, 1987; Ozubko, Gopie, & MacLeod, 2012; Ozubko, Major, & MacLeod, 2014). This finding is consistent with the distinctiveness heuristic (Dodson & Schacter, 2001), insofar as participants may be reluctant to classify new words as aloud because new words lack a distinct aloudness record.

The Present Research

In short, the problem with using a standard within-subject versus between-subjects design comparison to assess costs and benefits in recognition is that independent FA rates cannot be obtained for the within-subject (i.e., mixed-list) conditions. This concern is, of course, relevant to any research that compares mixed-list versus pure-list effects in recognition; we simply illustrate it here via the costs and benefits of production. The solution that we propose is to use instead a mixed-list versus pure-list experimental design that permits independent FAs to be obtained for both of the mixed-list conditions. In this research, we explored two such designs (in Experiments 2 and 3) using the costs and benefits of production as the testing ground.

To begin, in Experiment 1, we used a mixed-list versus pure-list design that conceptually replicated Bodner and colleagues' (2014) recent research. To achieve a high level of experimental power, we used a blocked design, in which all participants studied a pure-aloud list, a pure-silent list, and a mixed list (presented in random order). A forced-choice recognition test immediately followed each study list. Then, in Experiments 2 and 3, we made two different modifications to the standard within versus between designs. Each of these alterations allowed us to obtain separate FAs in the mixed-aloud and mixed-silent conditions, which we contend improves the accuracy of these d' calculations relative to those obtained in Experiment 1. In Experiment 2, participants made modality attributions at test: Instead of the standard old/new recognition test, participants who studied a mixed list were asked to classify each test word as either aloud, silent, or new. These attributions allowed us to differentiate "aloud FAs" from "silent FAs." In Experiment 3, participants who studied a mixed list were given a "pure test list" that contained only aloud words or only silent words. This design essentially made production a within-subject variable at study and a between-subjects variable at test, and thereby providing separate FAs for the mixed-list case.

Thus, our three experiments used variations of a mixed-list versus pure-list design to examine costs and benefits in recognition. In terms of hit rates, we expected to find benefits of reading aloud in a mixed list across the three experiments, a prediction derived straightforwardly from the distinctiveness account (MacLeod et al., 2010). Words read aloud in a mixed list should benefit from distinctive processing relative to words read aloud in a pure list.

We also hypothesized that dissociating the mixed-list FA rates in Experiments 2 and 3 would result in lower FA rates in the mixed-aloud condition than in the mixed-silent condition in both of these experiments. Assuming that, at the time of test, individuals attempt to retrieve diagnostic information about whether a word was studied aloud (Dodson & Schacter, 2001), then they should be relatively unlikely to mistake new words as having been studied aloud because new words lack this pertinent information. Moreover, participants may be less likely to FA aloud to new words following a mixed versus pure study list because aloud information stands out as distinct following the mixed list, which may increase the likelihood of participants using a distinctiveness heuristic at test.

In sum, by dissociating mixed-list FA rates in Experiments 2 and 3, we expected to find a mixed-list production effect not only in terms of hits, but also in terms of FAs. Importantly, if FA rates in the mixed-aloud condition decreased in Experiments 2 and 3 relative to Experiment 1, then memory discrimination as indexed by d' should correspondingly increase in this condition. Thus, although we did not expect to find a mixed-aloud d' benefit in Experiment 1, we predicted that such a benefit would emerge in the two subsequent experiments.

In terms of cost, we expected to replicate Bodner and colleagues' (2014) meta-analytic finding of a mixed-list cost of reading silently. Assuming that participants use a distinctiveness heuristic following a mixed list, this heuristic would bias them toward classifying silent items as unstudied because those items lack a distinct aloudness record (cf. Huff, Bodner, & Fawcett, 2015), which would lower the hit rates of the mixed-silent condition.

Experiment 1: A Standard Mixed Versus Pure Design

Our goal was to assess the costs and benefits of mixed-list production in recognition. To do so, we adopted the Bodner et al. (2014) approach of using a mixed-list versus pure-list design comparison. The pure-list means for aloud and silent items served as baselines against which the mixed-list means were compared. In terms of hits, we expected to find a mixed-list benefit of reading aloud and a mixed-list cost of reading silently, which would replicate Bodner and colleagues' (2014) results. A benefit would be consistent with aloud words being distinctively processed in a mixed study list. A cost would suggest a possible downside of encoding distinctive information: Following a mixed study list, participants who use a distinctiveness heuristic at test may be biased toward labeling words that lack a distinct record of having been encoded aloud (in this case, silently read words) as new.

According to our view, the drawback of this approach is that separate FAs cannot be obtained for the mixed-aloud and mixed-silent conditions, constraining us to using the same overall mixed-list FA rate to estimate FA rates for both the mixed-aloud condition and the mixed-silent condition. We argue that this constraint undermines valid assessment of mixed-list costs and benefits in terms of FA rates (and, therefore, also in terms of d').

Method

Participants. A total of 134 undergraduate students from the University of Waterloo participated in exchange for course credit.

Stimuli. The word pool consisted of 240 words obtained from the MRC Psycholinguistic Database (http://websites.psychology

.uwa.edu.au/school/MRCDatabase/uwa_mrc.htm). All words had frequencies of greater than 30 per million (Thorndike & Lorge, 1944) and were 5–10 letters long.

Apparatus. Stimuli were presented and responses were collected using E-Prime software (Psychology Software Tools Inc., Pittsburgh, PA) displayed on a 17" LCD monitor.

Procedure. Participants studied three different lists of words—pure aloud, pure silent, and mixed—in a counterbalanced order. List assignment to condition was also counterbalanced. For each list, font color was used to indicate whether the stimuli should be read aloud (blue font) or silently (white font).[3] Each study list consisted of 40 words presented centrally in 16 pt Courier New lowercase font against a black background. Each word was presented for 3 s followed by a 500-ms interstimulus interval.

Each of the three study lists was immediately followed by an 80-item recognition test made up of all 40 items from the preceding study phase randomly intermixed with 40 new distractor items. During the test, all items were presented one at a time in yellow font against a black background. Participants made keypress responses to label each item as old (*m*) or new (*c*). Each test item remained on the screen until the participant responded.

Results

Overview. We used this within-subject blocked design including all three conditions (pure aloud, pure silent, and mixed) to capitalize on a high level of statistical power. Whereas the typical within-subject versus between-subjects design would calculate costs and benefits by comparing independent groups of participants, our blocked design allowed for more powerful within-subject analyses of costs and benefits in the same large sample of participants. A post hoc power analysis using the statistical software G*Power (Erdfelder, Faul, & Buchner, 1996) showed that our experimental design had high statistical power (0.93) to detect small effects (Cohen's $d = 0.30$).

Not surprisingly, participants' recognition performance tended to decline across successive blocks, likely due to proactive interference or fatigue (for analyses of order effects, see Appendix A). Therefore, in addition to the overall analysis incorporating the data of all three blocks that we report in this Results section, we conducted supplemental analyses restricted to each participant's first block of data, which could not be influenced by block-order effects. Doing so resulted in a within-subject versus between-subjects design equivalent to that used by Bodner and colleagues (2014, Experiment 1). Appendix A reports these first-block analyses. Encouragingly, analyses of the first-block data set yielded a pattern of results very consistent with the full data analyses that we report here. Table 1 displays the hits rates, FA rates, and discrimination (d') values for the mixed-list condition and the two pure-list conditions. The means for the first-block data are also included in Table 1.

Hit rates. To compare the mixed-list and pure-list production effects for hits, we conducted a two-way repeated measures analysis of variance (ANOVA) in which study modality (aloud vs. silent) and study list type (mixed vs. pure) were both within-subject factors. Not surprisingly, there was a significant main effect of Study Modality, $F(1, 133) = 178.09$, $MSE = 0.01$, $p < .001$, $\eta^2 = 0.57$, indicating a robust overall production effect. The main effect of study list type was nonsignificant, $F(1, 133) = .29$, $MSE = 0.02$, $p = .59$, $\eta^2 = 0.002$.

Table 1
Experiment 1: Means (With SEs) for Each Group and Item Type

Measure/group	Aloud items	Silent items	New items
	Full		
Hits and false alarms			
Mixed-list group	.84 (.01)	.65 (.02)	.15 (.01)
Pure-list groups	.78 (.01)	.70 (.01)	.11 (.01)/.18 (.01)
Discrimination (d')			
Mixed-list group	2.27 (.07)	1.61 (.06)	
Pure-list groups	2.20 (.06)	1.66 (.07)	
	First block		
Hits and false alarms			
Mixed-list group	.87 (.02)	.66 (.03)	.12 (.01)
Pure-list groups	.82 (.02)	.73 (.02)	.11 (.01)/.16 (.02)
Discrimination (d')			
Mixed-list group	2.58 (.11)	1.80 (.10)	
Pure-list groups	2.35 (.10)	1.78 (.08)	

Note. Separate means were calculated for the full data and for the first-block data. The false alarm means for the pure-list groups refer to the aloud and silent groups, respectively.

More important, there was a significant Study Modality × Study List Type interaction, $F(1, 133) = 31.87$, $MSE = 0.01$, $p < .001$, $\eta^2 = 0.19$. Both the mixed-list production effect, $t(133) = 14.30$, $p < .001$, $d = 1.21$, and the pure-list production effect, $t(133) = 5.43$, $p < .01$, $d = 0.52$, were reliable. The significant Study Modality × Study List Type interaction signified that the mixed-list production effect was larger than the pure-list effect, consistent with the literature. This larger mixed-list production effect reflected both a significant benefit of reading aloud in a mixed list versus a pure aloud list, $t(133) = 4.96$, $p < .001$, $d = 0.44$, and a significant cost to reading silently in a mixed list versus a pure silent list, $t(133) = 2.92$, $p = .004$, $d = 0.28$. Both effect sizes were modest (in particular the cost).

FA rates. In these analyses, it is necessary to use participants' overall mixed-list FA rates as an estimate for both their mixed-aloud FA rates and their mixed-silent FA rates. This estimation of mixed-list FA rates allows the calculation of d' for both the mixed-aloud and mixed-silent conditions for use in the assessment of costs and benefits (see Bodner et al., 2014).[4]

We conducted another two-way repeated measures ANOVA to compare FA rates in the mixed-list and pure-list conditions. There

[3] Previous research (e.g., MacLeod et al., 2010; Bodner et al., 2014) has demonstrated that there is not a Font Color × Study Modality interaction, so we did not counterbalance these factors here.

[4] As argued throughout this article, this approach likely leads to an inaccurate estimation of mixed-aloud and mixed-silent FA rates. Thus, we recommend that memory researchers who use standard recognition tasks (like that used in our Experiment 1) not obtain separate mixed-aloud and mixed-silent FA rates in this manner and avoid calculating dprime. We have estimated separate mixed-list FA rates here to illustrate the potential inaccuracies of this approach, which we will then compare to the different approaches that we take in Experiments 2 and 3. Had we not separated mixed-list FA rates in Experiment 1, our FAs would be analyzed using a one-way repeated measures ANOVA, comparing the FA rates in the mixed, pure-aloud, and pure-silent conditions. This ANOVA yielded a significant main effect, $F(2, 166) = 20.24$, $MSE = 0.01$, $p < 0.001$, $\eta^2 = 0.13$.

was a significant main effect of Study Modality, $F(1, 133) = 33.27$, $MSE = 0.004$, $p < .001$, $\eta^2 = 0.20$, signifying that FA rates in the aloud conditions (i.e., mixed-aloud, pure-aloud) tended to be lower, overall, than FA rates in the silent conditions (i.e., mixed-silent, pure-silent). The main effect of Study List Type was nonsignificant, $F(1, 133) = 1.26$, $MSE = 0.01$, $p = .26$, $\eta^2 = 0.01$.

More important, there was a significant Study Modality × Study List Type interaction, $F(1, 133) = 33.27$, $MSE = 0.004$, $p < .001$, $\eta^2 = 0.20$.[5] This interaction reflected the fact that, in terms of FAs, the pure-list production effect was significantly larger than the mixed-list effect. This was not surprising. The pure-list production effect in terms of FAs was significant, $t(133) = 5.77$, $p < .001$, $d = 0.57$, consistent with participants using an aloudness distinctiveness heuristic to minimize FAs following a pure aloud study list (Dodson & Schacter, 2001). On the other hand, there was no difference between the mixed-aloud and mixed-silent FA rates, because a common value (the overall mixed-list FA rate) was used as the estimate for both conditions. The result of this shared mixed-list FA estimate was that the mixed-list production effect—which had been significantly larger than the pure-list effect in terms of hits—was now significantly smaller than the pure-list effect in terms of FAs.

The mixed-aloud FA rate was significantly greater than the pure-aloud FA rate, $t(133) = 4.29$, $p < .001$, $d = 0.42$, signifying an unexpected FA cost to reading aloud in a mixed study list. Thus, the mixed-list benefit of reading aloud in terms of hits was counteracted by a mixed-list cost of reading aloud in terms of FAs. The mixed-silent FA rate was significantly lower than the pure-silent FA rate, $t(133) = 2.34$, $p = .02$, $d = 0.18$, signifying an unexpected FA benefit to reading silently in a mixed study list. Thus, the mixed-list cost of reading silently in terms of hits was counteracted by a mixed-list benefit of reading silently in terms of FAs. Overall, then, the mixed-list versus pure-list effects for FA rates were opposite to the effects for hit rates—and this was the case for both aloud and silent words.

Memory discrimination (d')[6]. Last we conducted a two-way repeated measures ANOVA (parallel to those reported above), with d' as the dependent measure. Once again, there was a significant main effect of study modality, $F(1, 133) = 153.39$, $MSE = 0.31$, $p < .001$, $\eta^2 = 0.54$, indicating that memory discrimination was superior, overall, for words studied aloud versus silently. The main effect of study list type, however, was nonsignificant, $F(1, 133) = 0.03$, $MSE = 0.46$, $p = .87$, $\eta^2 = 0.00$, as was the Study Modality × Study List Type interaction, $F(1, 133) = 2.38$, $MSE = 0.23$, $p = .13$, $\eta^2 = 0.02$. This nonsignificant interaction suggests that the mixed-list d' production effect, $t(133) = 14.47$, $p < .001$, $d = 0.92$, did not differ reliably in magnitude from the pure-list d' production effect, $t(133) = 6.86$, $p < .001$, $d = 0.74$. The robust pure-list production effect in d' replicated the pure-list production effect reported by Fawcett (2013) and Bodner et al. (2014).

Because it was not possible to obtain separate FAs for the mixed-aloud condition and the mixed-silent condition, we argue that it therefore also is not possible to obtain precise calculations of memory discrimination (d') for these two conditions. Nonetheless, for the sake of comparison with our subsequent experiments, we calculated these estimates and show them in Table 1. Essentially, the mixed-list production effect in d' is a diluted version of the mixed-list production effect for hit rates (Cohen's $d = 1.21$, as shown above) because the same (constant) FA rate was applied both to the mixed-aloud condition and to the mixed-silent condition.

Comparing the mixed-aloud and pure-aloud conditions in terms of d' did not reveal a significant mixed-list benefit, $t(133) = 1.06$, *ns*. This d' result should be interpreted cautiously, though, because the mixed-aloud d' is derived from the estimated mixed-aloud FA rate. Comparing the mixed-silent and pure-silent conditions in terms of d' also did not reveal a significant mixed-list cost, $t(133) = 0.74$, *ns*. As with the d' benefit analysis, though, this d' cost analysis comes with the large caveat that the measure of d' for the mixed-silent condition is only an estimate. We cannot know the proportion of new words that participants in the mixed-list condition thought that they had studied silently (or aloud). Following this logic, we argue that the assessment of mixed-list costs in d' is also problematic.

Discussion

In Experiment 1, we used a "traditional" mixed-list versus pure-list design to examine the costs and benefits of recognition, replicating the recent research of Bodner and colleagues (2014, Experiment 1) and following the standard procedure for comparing within-subject to between-subjects designs. We found both mixed-list and pure-list production effects in terms of hits, as well as both a mixed-list benefit (of reading aloud) and a mixed-list cost (of reading silently).[7] In terms of memory discrimination (d'), we also found significant mixed-list and pure-list production effects, although we did not find evidence of costs or benefits. Thus, our results were mostly consistent with those of Bodner and colleagues (2014, Experiment 1), with the exception that their data suggested a mixed-list cost in d' (at $p = .07$), a pattern that we did not replicate despite our large-sample repeated measures design.

As we have argued, a shortcoming of the design used in Experiment 1—and common in the literature—is that separate FAs could not be obtained for the within-subject conditions, in this case for the aloud and silent mixed-list conditions. When conducted in the standard way, the recognition test yields only an overall mixed-list FA rate. Thus, d' could not be accurately measured for the mixed-aloud and mixed-silent conditions. Because the mixed-aloud and mixed-silent d' values are (potentially biased) estimates, the mixed-list d' production effect—as well as the d' benefit and d' cost—should be interpreted with caution.

[5] Note that this interaction was (necessarily) equivalent to the Study Modality main effect. This occurred because the same mixed-list FA rate was used as an estimate for the FA rates in both the mixed-aloud condition and the mixed-silent condition. Thus, the Study Modality main effect was completely driven by the simple effect of participants having fewer FAs in the pure-aloud condition than in the pure-silent condition.

[6] For computing d', ceiling hit rates and floor FA rates were adjusted using the 1/2N correction recommended by Macmillan and Creelman (2004). This correction was applied to 18 aloud hit rates, one silent hit rate, and 12 FA rates in the mixed condition; two hit rates and 10 FA rates in the pure-aloud condition; and zero hit rates and nine FA rates in the pure-silent condition.

[7] Although Bodner et al. (2014, Experiment 1) did not report hits in their article, analyzing their data revealed significant costs and benefits (both *p*s <.01). Their mean hit rates for each condition were: mixed aloud (M =0.83, SE = 0.02), mixed silent (M = 0.63, SE = 0.03), pure aloud (M = 0.76, SE = 0.02), and pure silent (M = 0.72, SE = 0.02).

In the following two experiments, to remedy this FA problem, we took two different approaches to modifying the standard within-subject versus between-subjects design. Each experiment represented a different methodological approach for obtaining separate FA rates for the two within-subject conditions (i.e., the mixed-aloud and the mixed-silent conditions), which we maintain leads to more accurate assessments of costs and benefits in d'.

Experiment 2: Modality Attributions

In Experiment 2, we altered the standard mixed-list versus pure-list design by asking participants to make modality attributions on the recognition test. On the recognition test following the mixed list, participants indicated whether each word was aloud, silent, or new. Correspondingly, participants chose between aloud and new following the pure-aloud list, and between silent and new following the pure-silent list. These modality attributions allowed us to obtain separate FAs in the mixed-aloud and mixed-silent conditions, which were then used to calculate condition-specific d' scores.

In line with the speech distinctiveness heuristic (Dodson & Schacter, 2001), we predicted that participants would be less likely to FA aloud than silent to new words following a mixed study list. Furthermore, we expected participants to show a mixed-list benefit of reading aloud because a mixed study list would encourage participants to process aloud words distinctively at study—which could, in turn, increase the likelihood of them using a distinctiveness heuristic at test. This enhanced distinctive processing afforded by studying words aloud in a mixed study list might have the twofold benefit both of boosting hit rates and of lowering FA rates relative to studying words aloud in a pure list.

Experiment 1 showed evidence of a mixed-list cost in terms of hits, consistent with previous research (Bodner et al., 2014; Hopkins & Edwards, 1972), so we expected to replicate this result in Experiment 2. By dissociating mixed-list FA rates in Experiment 2, we were also able to explore whether this cost extends to higher FA rates in the mixed-silent condition versus the pure-silent condition. That is, we could examine whether participants were more likely to FA silent to new words after studying a mixed list than after studying a pure-silent list.

Method

Participants. A total of 134 undergraduate participants from the University of Waterloo completed this experiment in exchange for course credit.

Stimuli and apparatus. These were the same as in Experiment 1.

Procedure. As in Experiment 1, we again used a blocked design in which participants studied three different lists of words—pure aloud, pure silent, and mixed—in a counterbalanced order. The study phase was identical to that of Experiment 1. The test phase was also largely the same, except for one important difference: Participants made modality attributions instead of old/new attributions on the recognition test. Participants made keypress responses to label each item as aloud (a), silent (s), or new (n). Following the mixed study list, participants had all three of these options whereas only the two relevant options were available following each of the pure study lists.

Results

Overview. As was the case with Experiment 1, preliminary analyses revealed the presence of block order effects—namely, participants' memory performance decreased across blocks (consistent with proactive interference or fatigue). The interested reader will find analyses of block order effects and of just the first-block data in Appendix B.

Importantly, the modality attributions in Experiment 2 not only allowed us to dissociate FA rates in the mixed-aloud and mixed-silent conditions, they also enabled us to fully dissociate hit rates in the mixed-aloud and mixed-silent conditions. Note that the standard old/new recognition memory test (like that used in Experiment 1) collapses mixed-aloud and mixed-silent hit rates in the sense that participants need only respond "old" to a word that they studied aloud or silently to be credited with a hit. Experiment 2, on the other hand, allowed us to measure hits more conservatively—as only occurring when the modality attribution was correct. This approach to scoring hits did, however, present problems with respect to assessing mixed-list costs and benefits. Namely: The mixed-list recognition task was different—and more difficult—than the pure-list recognition task. In the mixed-list recognition test, participants had to remember the study modality (aloud vs. silent), unlike in the pure-list recognition tests, in which they had to remember only whether the word had been studied.

Thus, to make the mixed-list data more comparable to the pure-list data in Experiment 2, we took a more lenient scoring approach. We collapsed the mixed-list hit rate data such that both aloud and silent responses were coded as old responses. Both of these responses were scored as hits as long as the word was studied (regardless of that word's actual study modality). We report these analyses first—examining mixed-list and pure-list production effects, as wells as costs and benefits—in the same manner as in Experiment 1. We then move on to the analyses of our modality data, in which only correct modality attributions were scored as hits, in addition to there being separable FA rates.

Table 2 displays the collapsed hit rates for aloud and silent items (for which aloud and silent responses were coded as old), and also displays FA rates and d'. As with Experiment 1, we display the means for both the full data and the first-block data. Note that, in contrast to Experiment 1, separate FAs were obtained for the mixed-aloud and mixed-silent conditions.

Hit rates. We conducted a two-way repeated-measures ANOVA on hit rates in which study modality (aloud vs. silent) and study list type (mixed vs. pure) were both within-subject factors. As expected, there was a robust main effect of study modality, $F(1, 133) = 118.74$, $MSE = 0.01$, $p < .001$, $\eta^2 = 0.47$, indicating an overall production effect. The main effect of study list type was nonsignificant, $F(1, 133) = 0.29$, $MSE = 0.02$, $p = .31$, $\eta^2 = 0.01$.

More important, there was a reliable Study Modality × Study List Type interaction, $F(1, 133) = 15.99$, $MSE = 0.01$, $p < .001$, $\eta^2 = 0.11$. Both the mixed-list production effect, $t(133) = 11.97$, $p < .001$, $d = 1.04$, and the pure-list production effect, $t(133) = 4.83$, $p < .001$, $d = 0.49$, were reliable. The significant Study Modality × Study List Type interaction signified that the mixed-list production effect was, as usual, larger than the pure-list effect. This larger mixed-list production effect reflected a significant benefit of reading aloud in a mixed-list versus a pure-list, $t(133) = 3.60$, $p < .001$, $d = 0.35$. There was not, however, a statistically

Table 2
Experiment 2: Means (With SEs) for Each Group and Item Type

Measure/group	Aloud items	Silent items	New items
	Full		
Hits and false alarms			
Mixed-list group	.83 (.01)	.68 (.01)	.04 (.01)/.16 (.01)
Pure-list groups	.78 (.01)	.70 (.01)	.09 (.01)/.17 (.01)
Discrimination (d')			
Mixed-list group	2.97 (.06)	1.65 (.06)	
Pure-list groups	2.35 (.06)	1.71 (.08)	
	First block		
Hits and false alarms			
Mixed-list group	.87 (.02)	.71 (.03)	.03 (.01)/.17 (.02)
Pure-list groups	.82 (.01)	.75 (.02)	.10 (.01)/.13 (.02)
Discrimination (d')			
Mixed-list group	3.20 (.10)	1.73 (.10)	
Pure-list groups	2.45 (.10)	2.06 (.13)	

Note. Aloud and silent responses were scored as hits if the word had been studied (even if the modality attribution was incorrect). Separate means were calculated for the full data and for the first-block data. The false alarm means for the pure-list groups refer to the aloud and silent groups, respectively.

significant cost of reading silently in a mixed-list versus a pure-list, although there was a trend in that direction that had a small effect size, $t(133) = 1.60$, $p = .11$, $d = 0.17$.

FA rates. We conducted another two-way repeated-measures ANOVA to compare FA rates in the mixed-list and pure-list conditions. There was a significant main effect of study modality, $F(1, 133) = 140.09$, $MSE = 0.01$, $p < .001$, $\eta^2 = 0.51$, signifying that FA rates tended to be lower, overall, in the aloud conditions (i.e., mixed-aloud, pure-aloud) than in the silent conditions (i.e., mixed-silent, pure-silent). The main effect of study list type was also significant, $F(1, 133) = 23.98$, $MSE = 0.01$, $p < .001$, $\eta^2 = 0.15$, reflecting that FA rates tended to be lower in the mixed-list conditions than in the pure-list conditions.

More important, there was a significant Study Modality × Study List Type interaction, $F(1, 133) = 9.79$, $MSE = 0.004$, $p = .002$, $\eta^2 = 0.07$. In keeping with participants using an aloudness distinctiveness heuristic following a pure aloud study list, participants had significantly lower FA rates in the pure-aloud condition than in the pure-silent condition, $t(133) = 11.34$, $p < .001$, $d = 1.28$. Participants also had significantly lower FA rates in the mixed-aloud condition than in the mixed-silent condition, $t(133) = 7.32$, $p < .001$, $d = 0.67$, consistent with participants using a distinctiveness heuristic following a mixed study list, as has previously been argued (e.g., Conway & Gathercole, 1987; MacLeod et al., 2010) The significant Study Modality × Study List Type interaction reflected the fact that, in terms of FAs, the mixed-list production effect was significantly larger than the pure-list effect.

This larger mixed-list production effect in terms of FAs can be attributed to a benefit of reading aloud in a mixed-list. That is, participants tended to FA aloud to new words less frequently following the mixed list than they did following the pure-aloud list, $t(133) = 8.78$, $p < .001$, $d = 0.74$. There was not, however, a statistically significant cost of reading silently in a mixed list in terms of FAs, $t(133) = 1.01$, $p = .31$, $d = 0.09$. Thus, the larger mixed-list versus pure-list FA effect reflected a FA benefit of reading aloud in a mixed list, without a corresponding cost of reading silently in a mixed list.

Memory discrimination (d')[8]. Last, we conducted a two-way repeated-measures ANOVA with d' as the dependent measure. There was a significant main effect of study modality, $F(1, 133) = 247.21$, $MSE = 0.47$, $p < .001$, $\eta^2 = 0.65$, signifying that memory discrimination was better, overall, for words studied aloud versus silently. The main effect of study list type was also reliable, $F(1, 133) = 29.45$, $MSE = 0.37$, $p < .001$, $\eta^2 = 0.18$, reflecting that memory discrimination was superior, overall, in the mixed-list conditions relative to the pure-list conditions.

More important, and consistent with our hypothesis, these main effects were qualified by a significant Study Modality × Study List Type interaction, $F(1, 133) = 41.33$, $MSE = 0.38$, $p < .001$, $\eta^2 = 0.24$. There was superior memory discrimination for words that were read aloud versus silently in a mixed list, $t(133) = 15.64$, $p < .001$, $d = 1.93$. There was also superior memory discrimination for words that were read aloud versus silently in pure lists, $t(133) = 8.06$, $p < .001$, $d = 0.79$. The significant interaction is consistent with the d' mixed-list production effect being reliably larger than the d' pure-list effect. This larger mixed-list production effect can be attributed to the d' benefit of reading aloud in a mixed list relative to a pure list, $t(133) = 9.72$, $p < .001$, $d = 0.89$. There was not, however, a reliable cost in d' of reading silently in a mixed list compared to a pure-silent list, $t(133) = 0.65$, $p = .52$, $d = 0.06$. (In this respect, the results for the full data set differed from the cost results in the first block data, which revealed significant, albeit modest [$d = 0.43$], cost in d'. We discuss this difference in Appendix B).

Modality attributions. In the analyses just reported, hits for the mixed-aloud and mixed-silent conditions were scored leniently in that participants were credited with a hit even when they chose the wrong study modality. In the present set of analyses, we applied the more stringent criterion of only crediting participants with a hit when they had indicated the correct study modality for an item. Table 3 shows participants' modality attributions for each type of word (i.e., aloud, silent, or new) that appeared on the recognition test following the mixed-study list. Note that this more stringent coding for hits only affected our mixed-list data. The pure-list data were not influenced because participants did not have multiple study modalities to choose from. Thus, the recognition test was more difficult following the mixed list because participants had three options to choose from rather than the two choices in the case of the pure lists.

Because of this difference in recognition task difficulty, an assessment of costs and benefits was bound to be skewed in the

[8] For computing d', ceiling hit rates and floor FA rates were adjusted using Macmillan and Creelman's (2004) 1/2N correction. This correction was applied to 14 aloud hit rates and 68 FA rates in the mixed-aloud condition; zero hit rates and 10 FA rates in the mixed-silent condition; one hit rate and 21 FA rates in the pure-aloud condition; and two hit rates and 12 FA rates in the pure-silent condition. Notably, slightly more than half of the participants (68/134) were at the floor in terms of their FA rates in the mixed-aloud condition (consistent with the robust FA benefit in the mixed-aloud condition). This large number of FA adjustments arguably yielded a conservative calculation of d' in the mixed-aloud condition relative to the other conditions of this experiment, which did not have as large of a disparity between ceiling hit rates and floor FA rates. Nevertheless, a robust d' benefit was observed.

Table 3
Experiment 2: Means (With SEs) of Participants' Modality Attributions for Each Type of Word (Aloud, Silent, or New) That Appeared on the Recognition Test That Followed the Mixed Study List

Actual modality/ attributed modality	Aloud items	Silent items	New items
	Full		
Aloud (attributed)	.58 (.02)	.15 (.01)	.04 (.01)
Silent (attributed)	.25 (.01)	.52 (.01)	.16 (.01)
New (attributed)	.17 (.01)	.32 (.01)	.80 (.01)
	First block		
Aloud (attributed)	.64 (.03)	.15 (.02)	.03 (.01)
Silent (attributed)	.23 (.02)	.57 (.02)	.17 (.02)
New (attributed)	.13 (.02)	.29 (.02)	.80 (.02)

Note. Separate means were calculated for the full data and for the first-block data.

direction of costs. Predictably, then, when only correct modality attributions were scored as hits, the mixed-list cost of reading silently was exacerbated. The mixed-silent hit rate ($M = 0.52$, $SE = 0.01$) was significantly lower than the pure-silent hit rate ($M = 0.70$, $SE = 0.01$), $t(133) = 9.71$, $p < .001$, $d = 1.09$, and the mixed-silent d' ($M = 1.21$, $SE = 0.06$) was significantly lower than the pure-silent d' ($M = 1.71$, $SE = 0.08$), $t(133) = 6.04$, $p < .001$, $d = 0.62$.

Moreover, there now was a cost of reading aloud in the mixed list, in contrast to the benefit reported earlier. The mixed-aloud hit rate ($M = 0.58$, $SE = 0.02$) was significantly lower than the pure-aloud hit rate ($M = 0.78$, $SE = 0.01$), $t(133) = 11.45$, $p < .001$, $d = 1.17$, and the mixed-aloud d' ($M = 2.15$, $SE = 0.06$) was significantly lower than the pure-aloud d' ($M = 2.35$, $SE = 0.06$), $t(133) = 3.12$, $p = .002$, $d = 0.28$. This mixed-list cost is not surprising given the greater difficulty of the mixed-list recognition task. Despite this shift toward cost, though, there still was a mixed-list production effect in terms of both hits, $t(133) = 2.87$, $p = .005$, $d = 0.30$, and d', $t(133) = 11.51$, $p < .001$, $d = 1.32$.

Although the mixed-list modality data may be problematic with respect to assessing costs and benefits, they have provided an informative test of the distinctiveness account (MacLeod et al., 2010). First, there is evidence that participants tend to conflate silent and new words at test. As reported earlier in this Results section, participants were more likely to FA silent than aloud to new words, $t(133) = 11.34$, $p < .001$, $d = 1.28$. Participants were also more likely to miscategorize silent words as new than as aloud, $t(133) = 8.93$, $p < .001$, $d = 1.24$. Both of these results are consistent with participants using a distinctiveness heuristic at test, as both silent and new words lack a distinct record of speech. There was, however, a result that was not in line with the distinctiveness heuristic: Participants were more likely to misclassify aloud words as silent than as new, $t(133) = 4.06$, $p < .001$, $d = 0.52$. If participants were strictly relying on a distinctiveness heuristic, they should have been equally likely to misjudge aloud words as being silent versus new because both types of words lack distinct aloud information.[9]

Thus, our modality results were not entirely consistent with the distinctiveness account. Arguably, these results are more consistent with the evaluated-strength account, which Bodner and Taikh (2012) offered as an alternate explanation of the production effect. According to those authors, participants may consciously judge the signal strength (Wickelgren, 1969) of items on a recognition test. When a word has a particularly strong (i.e., familiar) record at test, participants are biased toward making the attribution that they studied the word aloud. Bodner and Taikh found evidence that this attributional bias influenced participants' source judgments (in the context of a list discrimination task). The same biases may have influenced participants' source judgments in the recognition test used here.

Our modality results were consistent with the evaluated-strength account as follows. First, assuming that words read silently had lower signal strength than those read aloud, participants may have been biased to FA silent rather than aloud to new words because silent words were more similar to new words in terms of strength/familiarity. Second, the finding that participants were more likely to misattribute silent words as new than as aloud could have occurred because silent words were more similar in signal strength to new words than to aloud words. Third, in the same vein, participants may have been biased to misattribute aloud words as silent rather than as new due to aloud and silent words having signal strength more similar to each other than to new words.

These assumptions regarding signal strength cannot be tested with the present data. Nonetheless, the pattern of modality results suggests that participants may not have been relying entirely on distinctiveness information when making their modality attributions at test (otherwise, they probably would not have shown the attributional bias of categorizing aloud words as silent rather than as new). The present results suggest that participants may use both distinctiveness and strength information diagnostically at test; this possibility is consistent with Ozubko, Gopie, and MacLeod's (2012) finding that the advantage of production results from a combination of greater recollection and greater familiarity for the aloud versus silent items.

Comparing the mixed-list benefit in Experiments 1 and 2. In Experiment 2, we were able to obtain separate FA rates in the mixed-aloud and mixed-silent conditions (unlike in Experiment 1, in which the overall mixed-list FA rate was used as an estimate for both of these values). This design change resulted in a significantly lower mixed-aloud FA rate than mixed-silent FA rate in Experiment 2. This difference supports our contention that using a single FA rate for both of the mixed-list conditions misrepresents what is actually occurring in memory. This result is also consistent with participants using a speech distinctiveness heuristic (Dodson & Schacter, 2001) to avoid false alarming aloud to new words.

In addition, we predicted that dissociating mixed-list FA rates in Experiment 2 would result in a lower FA rate—and, consequently, better memory discrimination—in the mixed-aloud condition of Experiment 2 relative to the mixed-aloud condition of Experiment 1, in which mixed-list FA rates were not dissociated. Consequently, we expected that comparing the mean d' values across experiments would reveal a larger mixed-list benefit of reading aloud in Experiment 2 versus Experiment 1. We also examined whether this design change resulted in a larger mixed-list cost to reading silently in Experiment 2.

[9] We thank Glen Bodner for raising this point as a reviewer.

To test whether the benefit of reading aloud was larger in Experiment 2 than in Experiment 1, we compared recognition performance (for hits, FAs, and d') across these two experiments by running two-way ANOVAs with Study list type (mixed-aloud vs. pure-aloud) as a within-subject factor and experiment (Experiment 1 vs. Experiment 2) as a between-subjects factor. For Experiment 2, we used the hit rates obtained from the lenient scoring approach (which scored both aloud and silent modality attributions as hits as long as the word had been studied). Analyses that used the stringent approach (which scored only correct modality attributions as hits) are reported in footnotes.

For hit rates, the two-way ANOVA revealed a significant main effect of study list type, $F(1, 266) = 35.35$, $MSE = 0.01$, $p < .001$, $\eta^2 = 0.12$, reflecting an overall mixed-list benefit of reading aloud in terms of hits. The main effect of experiment, however, was not significant nor was the Experiment × Study List Type interaction (both Fs < 1), suggesting that the two experiments yielded comparable mixed-list benefits in terms of hits (when the lenient scoring approach was used in Experiment 2).[10]

Next, we ran this ANOVA with FA rates as the dependent measure. The main effect of study list type was not reliable, $F(1, 174) = 1.10$, $MSE = 0.01$, $p = .30$, $\eta^2 = 0.004$. The main effect of experiment, however, was significant, $F(1, 174) = 6.16$, $MSE = 0.01$, $p = .01$, $\eta^2 = 0.03$, signifying overall lower FA rates in Experiment 2 versus Experiment 1, $F(1, 266) = 57.36$, $MSE = 0.01$, $p < .001$, $\eta^2 = 0.18$. More important, and of main interest, the Experiment × Study List Type interaction was again significant, $F(1, 266) = 69.01$, $MSE = 0.01$, $p < .001$, $\eta^2 = 0.21$, indicating a larger mixed-list benefit of reading aloud in Experiment 2 than in Experiment 1. FA rates in the mixed-aloud condition were lower in Experiment 2 than in Experiment 1, $t(193.28) = 10.57$, $p < .001$, $d = 1.29$. FA rates were also significantly lower in the pure-aloud condition in Experiment 2 than in Experiment 1, $t(266) = 1.98$, $p = .05$, $d = 0.24$, a small effect. Although participants tended, overall, to have lower FA rates in Experiment 2 than in Experiment 1, this effect clearly was more pronounced in the mixed-aloud condition than in the pure-aloud condition, which resulted in the significant interaction.

Last, we ran this ANOVA with d' as the dependent measure. An identical pattern of results was observed as reported above with FAs. The main effect of Study List Type was significant, $F(1, 266) = 54.46$, $MSE = 0.30$, $p < .001$, $\eta^2 = 0.17$, signifying an overall mixed-list benefit of reading aloud. The main effect of experiment was also significant, $F(1, 266) = 35.35$, $MSE = 0.71$, $p < .001$, $\eta^2 = 0.12$, reflecting overall higher d' values in Experiment 2. Most important, the Experiment × Study List Type interaction was significant, $F(1, 266) = 35.90$, $MSE = 0.30$, $p < .001$, $\eta^2 = 0.11$, indicating a larger mixed-list benefit of reading aloud in Experiment 2 than in Experiment 1. Participants' average mixed-aloud d' was higher in Experiment 2 than in Experiment 1, $t(266) = 7.97$, $p < .001$, $d = 0.98$. Participants' average pure-aloud d' was also slightly higher in Experiment 2 than it was in Experiment 1, $t(266) = 1.83$, $p = .07$, $d = 0.22$. Thus, it appears that the design change implemented in Experiment 2 mainly improved memory for mixed-aloud items while having only a small (nonsignificant) effect on their memory for pure-aloud items.[11]

Comparing the mixed-list cost in Experiments 1 and 2. The preceding results showed that participants had a larger mixed-list benefit of reading aloud in Experiment 2 than they did in Experiment 1 (when hit rates were scored leniently). As predicted, this increased d' benefit seemed to arise from a decrease in the FA rate in the mixed-aloud condition in Experiment 2 relative to Experiment 1. Next, we examined whether the mixed-list cost of reading silently differed between these two experiments. We conducted a series of two-way ANOVAs, parallel to those above, with study list type (mixed-silent vs. pure-silent) as a within-subject factor and experiment (Experiment 1 vs. Experiment 2) as a between-subjects factor.

First, we ran this ANOVA with hits (leniently scored) as the dependent factor. The ANOVA revealed a significant main effect of study list type, $F(1, 266) = 10.26$, $MSE = 0.02$, $p = .002$, $\eta^2 = 0.04$, signifying an overall mixed-list cost of reading silently across both experiments. Both the main effect of experiment and the Experiment × Study List Type interaction were nonsignificant (both Fs < 1), suggesting that the size of the mixed-list cost did not differ between experiments (when the lenient scoring approach was used in Experiment 2).[12]

In terms of FAs, a two-way ANOVA revealed, unexpectedly, that participants tended to make significantly fewer FAs in the mixed-silent condition than in the pure-silent condition, $F(1, 266) = 5.20$, $MSE = 0.01$, $p = .02$, $\eta^2 = 0.02$. This mean difference was quite small (2%). More importantly, neither the main effect of experiment nor the Experiment × Study List Type interaction was statistically significant (both Fs < 1), indicating that the patterns of FAs in the mixed-silent and pure-silent conditions were very similar over these two experiments.

[10] When we stringently scored hit rates in Experiment 2, the two-way ANOVA revealed a significant main effect of study list type, $F(1, 266) = 43.29$, $MSE = 0.01$, $p < 0.001$, $\eta^2 = 0.14$, signifying that mixed-aloud hit rates were lower overall than pure-aloud hit rates (in contrast to the overall mixed-list benefit that arose when hits were leniently scored). The main effect of experiment was also significant, $F(1, 266) = 71.46$, $MSE = 0.03$, $p < 0.001$, $\eta^2 = 0.21$, indicating that aloud hit rates were lower, overall, in Experiment 2 than in Experiment 1. The Experiment × Study List Type interaction was also robust, $F(1, 266) = 149.56$, $MSE = 0.01$, $p < 0.001$, $\eta^2 = 0.36$, reflecting the fact that there was a mixed-list benefit to reading aloud in Experiment 1, $t(133) = 4.96$, $p < 0.001$, $d = 0.44$, but a significant mixed-list cost to reading aloud in Experiment 2, $t(133) = 11.45$, $p < 0.001$, $d = 1.17$. We have already argued that hit rates in the mixed-aloud condition of Experiment 2 were particularly low due to the added difficulty of making correct modality attributions.

[11] When stringently scored hit rates were used to derive the mixed-aloud d' values in Experiment 2, the ANOVA yielded nonsignificant main effects both for Study List Type, $F(1, 266) = 1.87$, $MSE = 0.31$, $p = 0.17$, $\eta^2 = 0.01$, and for Experiment ($F < 1$). The Experiment × Study List Type interaction, however, was significant, $F(1, 266) = 8.48$, $MSE = 0.31$, $p = 0.004$, $\eta^2 = 0.03$, reflecting the fact that mixed-aloud and pure-aloud d' values differed significantly in Experiment 2, $t(133) = 3.12$, $p = 0.002$, $d = 0.28$, but not in Experiment 1, $t(133) = 1.06$, *ns*.

[12] When hits were scored stringently in Experiment 2, the ANOVA also revealed a significant main effect of Study List Type, $F(1, 266) = 84.31$, $MSE = 0.02$, $p < 0.001$, $\eta^2 = 0.24$, signifying a robust overall mixed-list cost of reading silently. The main effect of experiment was also significant, $F(1, 266) = 14.56$, $MSE = 0.04$, $p < 0.001$, $\eta^2 = 0.05$, indicating that silent hit rates were lower, overall, in Experiment 2 than in Experiment 1. The Experiment × Study List Type interaction was also significant, $F(1, 266) = 27.90$, $MSE = 0.02$, $p < 0.001$, $\eta^2 = 0.09$, reflecting the fact that the mixed-list cost of reading silently was larger in Experiment 2, $t(133) = 9.71$, $p < 0.001$, $d = 1.09$, than it was in Experiment 1, $t(133) = 2.92$, $p = 0.004$, $d = 0.28$. We contend that hit rates in the mixed-silent condition of Experiment 2 were particularly low due to the increased difficulty of making correct modality attributions.

Last, we ran this ANOVA with d' as the dependent measure. This two-way ANOVA yielded nonsignificant main effects of Study List Type and Experiment, as well as a nonsignificant Experiment × Study List Type interaction (all Fs < 1). Thus, both experiments yielded an equivalently small, nonsignificant cost of reading silently in a mixed list.[13]

In summary, these combined analyses did not provide evidence that the design change implemented in Experiment 2 (i.e., dissociating FA rates through modality attributions) resulted in a greater mixed-list cost of reading silently. Indeed, although the mixed-list cost of reading silently was reliable (but small) in both experiments in terms of hits, it was not reliable in terms of d'. The evidence was much stronger that the mixed-list versus pure-list design implemented in Experiment 2 revealed an increased mixed-list *benefit* of reading aloud relative to Experiment 1. However, the d' results hold only when the hit rates in Experiment 2 were scored leniently (with the purpose of equating the difficulty of the recognition tasks across the two experiments). A stringent scoring of hit rates revealed stronger d' costs of reading silently in Experiment 2 than in Experiment 1, as well as stronger d' costs of reading aloud. But regardless of how hit rates were scored in Experiment 2, it is clear that dissociating FA rates had the anticipated effect of decreasing FA rates in the mixed-aloud condition relative to the mixed-aloud FA estimate obtained in Experiment 1 (in which mixed-list FA rates could not be dissociated).

Discussion

In Experiment 2, we used a mixed-list versus pure-list design in which participants made modality attributions on the recognition test. Unlike Experiment 1, in which we were constrained to use the overall mixed-list FA rate as an estimate for both the mixed-aloud condition and the mixed-silent condition, the modality attributions in Experiment 2 allowed us to obtain separate FA rates in these conditions. We contend that these dissociated FA rates led to more accurate calculations of d' in the mixed conditions than were obtained using a collapsed FA rate (Experiment 1), in turn leading to more meaningful assessments of mixed-list costs and benefits.

A drawback of this design, however, was that the mixed-list recognition test (which had three response options) was more difficult than the pure-list recognition tests (which had only two response options). Thus, when only correct modality attributions were scored as hits in the mixed list, the results showed a strong memorial cost of studying words in a mixed list. To draw more meaningful comparisons between the mixed-list and the pure-list recognition tests, we adopted the more lenient criterion of binning aloud and silent responses in the mixed-list recognition task as old, and scoring them as hits as long as the word had been studied. This approach revealed significant benefits of mixed-list production.

Importantly, dissociating FA rates in Experiment 2 revealed that participants were substantially less likely to FA aloud than silent to new words. This result is consistent with participants using a distinctiveness heuristic: They may have refrained from classifying new items as aloud because new items lack distinct auditory and articulatory records. Comparing the results of Experiment 1 and 2 highlighted the fact that participants' hit rates were quite stable across these two designs—showing significant costs and benefits. In terms of FA rates, the mixed-aloud FA rate in Experiment 2 was much lower than the estimated mixed-aloud FA rate used in Experiment 1. This difference led to a mixed-list d' benefit in Experiment 2 that was not evident in Experiment 1. Conversely, there was no evidence of a mixed-list cost of reading silently in the full data of Experiment 2 (although a small cost was found in an analysis of the first-block data, reported in Appendix B).

In sum, having participants make modality attributions at test in Experiment 2 yielded a robust mixed-list d' benefit of reading aloud, a novel finding. This mixed-list benefit relative to reading aloud in a pure list was evident in terms of both higher hit rates and lower FA rates. Conversely, the mixed-list cost of reading silently (demonstrated in Bodner et al., 2014) was not amplified by this modality attribution manipulation. Memory for mixed-silent items was no worse in Experiment 2 than in Experiment 1.

This evidence of a mixed-list d' benefit only emerged, however, when a lenient scoring criterion was used for the recognition task following the mixed list, such that both aloud and silent responses were scored as hits for all studied words, regardless of modality. A more stringent criterion of scoring only correct modality attributions as hits yielded evidence of mixed-list d' costs—both for reading silently and for reading aloud. But regardless of how hits were scored, Experiment 2 revealed that participants seldom false alarmed aloud to new words after studying a mixed list. Mixed-aloud FA rates were lower than mixed-silent FA rates—consistent with the distinctiveness heuristic—and were also lower than pure-aloud FA rates.

Experiment 3: Pure Tests

In Experiment 2, we found a robust benefit of reading aloud in a mixed list when participants made modality attributions at test (which dissociated the mixed-list FA rates). Assessing costs and benefits using this design was not ideal, however, because participants had three modality choices after studying a mixed list, but only two choices after studying a pure list. In Experiment 3, therefore, to address this issue, we took a different approach to dissociating mixed-list FA rates. We used a mixed-list versus pure-list design in which participants who studied a mixed list received only one type of studied item at test—either aloud or silent. Thus, the recognition test in the mixed-aloud condition was identical in format to that in the pure-aloud condition, and the recognition test in the mixed-silent condition was identical in format to that in the pure-silent condition. In other words, the mixed-aloud and mixed-silent tests were run between-subjects, unlike the typical within-subject mixed test used in the previous experiments. This design allowed us to examine the influence of study list type (i.e., mixed vs. pure) on the costs and benefits of production, while controlling for the form of the recognition test.

[13] When stringently scored hit rates were used to derive the mixed-silent d' values in Experiment 2, the ANOVA yielded a significant main effect of study list type, $F(1, 266) = 24.92$, $MSE = 0.41$, $p < 0.001$, $\eta^2 = 0.09$, signifying that mixed-silent d' values tended to be lower, overall, than pure-silent d' values. The main effect of experiment was also significant, $F(1, 266) = 5.20$, $MSE = 0.77$, $p = 0.02$, $\eta^2 = 0.02$, indicating that silent d' values tended to be lower in Experiment 2 than in Experiment 1. Qualifying these main effect was a significant Experiment × Study List type interaction, $F(1, 266) = 8.48$, $MSE = 0.31$, $p = 0.004$, $\eta^2 = 0.03$, reflecting the fact that there was a significant mixed-list d' cost in Experiment 2, $t(133) = 6.04$, $p < 0.001$, $d = 0.62$, but not in Experiment 1, $t(133) = 0.74$, *ns.*

We hypothesized that using this "pure test" approach to dissociating mixed-list FA rates would yield significantly lower FA rates in the mixed-aloud condition than in the mixed-silent condition. This prediction again derived from the distinctiveness account. Participants in the mixed-aloud condition could use a speech distinctiveness heuristic to avoid false alarming aloud to new words, just as individuals have been shown to use this distinctiveness heuristic following pure-aloud lists (Dodson & Schacter, 2001). Indeed, participants may be more inclined to use a distinctiveness heuristic following a mixed study list, assuming that they have already distinctively processed aloud information when studying the mixed list, which may increase the salience of a distinctiveness heuristic.

The predictions regarding costs were less straightforward. The previous two experiments showed significant mixed-list costs in terms of hit rates, which we expected to replicate in Experiment 3. Dissociating the mixed-list FA rates in Experiment 3 allowed us to directly compare FA rates in the mixed-silent and pure-silent conditions. The results of Experiment 2, in which we also dissociated mixed-list FA rates, did not reveal consistent evidence that there was a mixed-list cost of reading silently, in terms of higher FA rates. Thus, we did not expect to find evidence of such a cost in Experiment 3.

In this experiment, we used the same materials as Bodner and colleagues (2014) had used in their Experiment 1. We predicted that dissociating the mixed-list FA rates would lead to a significantly larger d' benefit in this experiment than in their Experiment 1, in which FA rates were not dissociated. We expected that a mixed-list benefit in our experiment would be driven both by higher hit rates and by lower FA rates in our mixed-aloud condition versus our pure-aloud condition. Conversely, the size of the benefit in Experiment 1 of Bodner et al. (2014) may have been constrained by the fact that their design did not yield separate mixed-list FA rates.

Method

Participants. A total of 224 participants from the University of Waterloo participated in this experiment. Fifty-six participants took part in each of the four conditions: mixed-aloud, mixed-silent, pure-aloud, and pure-silent. As remuneration, all participants received credit toward one of their psychology courses.

Stimuli. The stimuli were comprised of the same 100 words that were used in Experiment 1 of Bodner et al. (2014) and in previous production research (e.g., MacLeod et al., 2010). Consistent with the previous two experiments, words were 5 to 10 letters long and had frequencies of greater than 30 per million (Thorndike & Lorge, 1944). Sixteen of the words in the word pool of Experiment 3 were also in the stimulus set used in Experiments 1 and 2. Words in this experiment were presented in counterbalanced font colors (green and orange) to indicate their study modality (aloud vs. silent).

Procedure. Participants completed a study phase similar to that in Experiment 1. They studied 50 words, each presented for 2 s and separated by a 500-ms interstimulus interval from the next one. Stimuli were presented either in green or in orange 36-pt Arial font against a white background. Next, participants completed a single forced-choice recognition test, using the "c" key to indicate that a word was judged to be studied, and the "m" key to indicate that a word was judged to be new. For the mixed lists, the tests consisted of the 25 items that had been studied aloud along with 25 distractors or of the 25 items that had been studied silently along with 25 distractors, resulting in "pure tests." Prior to the test, participants in the mixed-list conditions were informed that words studied in the other modality would not appear on that test. This instruction was not given to participants who studied a pure list (because they had only one study modality). In the Discussion, we consider the possible effect that this additional instruction may have had on the memory performance of participants in the mixed-aloud condition.

Both of the pure-list tests were constructed by randomly selecting 25 old words from the study phase to intermix with 25 distractors. Note that this procedure was identical to Experiment 1 of Bodner et al. (2014), except for the fact that our participants who studied a mixed list were only tested on one type of study item as opposed to both types, allowing us to obtain separate FA rates for aloud and silent items. It was also our intention to keep the total length of the test list the same for the pure lists as for the mixed lists—both contained 25 old words and 25 new words. This resulted in participants getting a 50-item recognition test, half the length of that used in Bodner and colleagues' (2014) Experiment 1.

Results

Table 4 displays the mean hits, FAs, and d' values for participants who studied either aloud or silent items in either a mixed list or a pure list. For comparison, we also include Bodner and colleagues' (2014) Experiment 1 results. Because our recognition test was half the length of that used by Bodner and colleagues, our participants, not surprisingly, showed superior overall performance.

Hit rates. We ran a two-way repeated-measures ANOVA on hit rates in which study modality (aloud vs. silent) and study list type (mixed vs. pure) were both between-subjects factors. There

Table 4
Means (With SEs) for Each Group and Item Type in Our Experiment 3 and in Experiment 1 of Bodner et al. (2014)

Measure/group	Aloud items	Silent items	New items
Experiment 3			
Hits and false alarms			
Mixed-list group	.88 (.01)	.73 (.02)	.07 (.01)/.17 (.02)
Pure-list groups	.84 (.02)	.79 (.02)	.11 (.01)/.18 (.02)
Discrimination (d′)			
Mixed-list group	2.81 (.09)	1.77 (.08)	
Pure-list groups	2.47 (.10)	1.96 (.10)	
Bodner et al. (2014, Experiment 1)			
Hits and false alarms			
Mixed-list group	.83 (.02)	.63 (.03)	.17 (.02)
Pure-list groups	.76 (.02)	.72 (.02)	.09 (.01)/.18 (.02)
Discrimination (d')			
Mixed-list group	2.18 (.12)	1.44 (.10)	
Pure-list groups	2.21 (.09)	1.72 (.12)	

Note. The false alarm means for the pure-list groups refer to the aloud and silent groups, respectively.

was once again a significant main effect of study modality, $F(1, 220) = 32.05$, $MSE = 0.02$, $p < .001$, $\eta^2 = 0.13$, indicating an overall production effect. The main effect of study list type was nonsignificant, $F(1, 220) = 0.65$, $MSE = 0.02$, $p = .42$, $\eta^2 = 0.003$.

Most important, there was a reliable Study Modality × Study List Type interaction, $F(1, 220) = 8.01$, $MSE = 0.02$, $p < .001$, $\eta^2 = 0.04$. Both the mixed-list production effect, $t(82.75) = 5.65$, $p < .001$, $d = 1.07$, and the pure-list production effect, $t(110) = 2.15$, $p < .05$, $d = 0.41$, were reliable. The significant Study Modality × Study List Type interaction signified that the mixed-list production effect was, as is typically found, larger than the pure-list effect. This larger mixed-list production effect reflected both a marginal benefit of reading aloud in a mixed list versus a pure list, $t(110) = 1.78$, $p = .08$, $d = 0.34$, and a cost to reading silently in a mixed-list versus a pure list, $t(99.68) = 2.22$, $p = .03$, $d = 0.42$. Both effect sizes were modest.

FA rates. We carried out another two-way repeated measures ANOVA to compare FA rates in the mixed-list and pure-list conditions. There was a significant main effect of study modality, $F(1, 220) = 28.66$, $MSE = 0.01$, $p < .001$, $\eta^2 = 0.12$, indicating that FA rates tended to be lower, overall, in the aloud conditions (i.e., mixed-aloud, pure-aloud) than in the silent conditions (i.e., mixed-silent, pure-silent). The main effect of study list type was nonsignificant, $F(1, 220) = 2.66$, $MSE = 0.01$, $p = .10$, $\eta^2 = 0.01$.

Of main interest, the Study Modality × Study List Type interaction was nonsignificant, $F(1, 220) = 0.51$, $MSE = 0.01$, $p = .48$, $\eta^2 = 0.002$. Consistent with the use of an aloudness distinctiveness heuristic, participants had significantly lower FA rates in the mixed-aloud condition than in the mixed-silent condition, $t(90.35) = 5.02$, $p < .001$, $d = 0.95$, and they also had significantly lower FA rates in the pure-aloud condition than in the pure-silent condition, $t(93.08) = 2.91$, $p < .01$, $d = 0.55$. The nonsignificant interaction suggests that these two FA production effects did not differ reliably in magnitude. There was nonetheless a significant benefit of reading aloud in a mixed list versus a pure list in terms of lower FA rates, $t(100.82) = 2.22$, $p = .03$, $d = 0.42$. There was not a statistically significant cost of reading silently in a mixed list in terms of FAs, $t(110) = 0.54$, $p = .59$, $d = 0.10$.

Memory discrimination (d')[14]. Last, we conducted a two-way repeated measures ANOVA on d'. There was a significant main effect of study modality, $F(1, 220) = 68.49$, $MSE = 0.49$, $p < .001$, $\eta^2 = 0.24$, indicating better overall memory discrimination for words read aloud versus silently. The main effect of study list type was not reliable, $F(1, 220) = 0.55$, $MSE = 0.49$, $p = .46$, $\eta^2 = 0.002$.

Importantly, and consistent with our predictions, there was a statistically significant Study Modality × Study List Type interaction, $F(1, 220) = 8.04$, $MSE = 0.49$, $p < .001$, $\eta^2 = 0.04$. There was superior memory discrimination for words that were read aloud versus silently in a mixed list, $t(110) = 8.69$, $p < .001$, $d = 1.66$. There was also superior memory discrimination for words that were read aloud versus silently in pure lists, $t(110) = 3.54$, $p < .001$, $d = 0.68$. The significant interaction is consistent with the d' mixed-list production effect being reliably larger than the d' pure-list effect. This larger mixed-list production effect reflects the hypothesized d' benefit of reading aloud in a mixed list relative to a pure list, $t(110) = 2.54$, $p = .01$, $d = 0.48$, consistent with the results of Experiment 2. There was not, however, a reliable cost in d' of reading silently in a mixed list compared to a pure list, again consistent with Experiment 2, although there was a trend in that direction that had a small effect size, $t(104.53) = 1.47$, $p = .14$, $d = 0.28$.

Discussion

In Experiment 3, we took a second approach to dissociating FAs at test following a mixed study list. Participants in the mixed-list groups were given a pure recognition test in which their recognition for only aloud or only silent items was tested. In all other respects, the design of this experiment was identical to Experiment 1 of Bodner et al. (2014). Memory for aloud and silent items studied in a pure list again served as a baseline against which benefits and costs in mixed lists were assessed.

Replicating the results of Experiments 1 and 2, we found both mixed-list and pure-list production effects. More important, we again found a significant benefit in memory discrimination for reading aloud in a mixed list relative to a pure list. This benefit was evident in terms of hits ($p = .08$), consistent with the results of the previous experiments in this article as well as with Experiment 1 of Bodner et al. (2014). There was also a significant mixed-list benefit in terms of a lower FA rate in the mixed-aloud condition versus the pure-aloud condition, as well as a mixed-list d' benefit.

In addition, we found a mixed-list cost of reading silently in terms of hits, again consistent with the results of the previous experiments in this article as well as with Experiment 1 of Bodner et al. (2014). There was not, however, a significant cost in terms of FAs, and the d' cost was trending, but not reliable. The cost of reading silently in a mixed list appeared to be roughly equivalent to the cost observed in these other two experiments, suggesting that this cost is not influenced in any meaningful way by dissociating mixed-list FA rates.

In summary, the results of Experiment 3 are largely consistent with those of Experiment 2, demonstrating superior memory discrimination for words studied aloud in a mixed list versus a pure list. It is worth emphasizing that the mixed-list d' benefit was significantly larger in the present Experiment 3 than in Experiment 1 of Bodner et al. (2014). (A thorough comparison of our Experiment 3 and their Experiment 1 is presented in Appendix C). Because our experiment was the same as Bodner et al.'s experiment in almost every respect—except for having a pure recognition test rather than a standard (mixed) recognition test—we assume that our pure-test manipulation was responsible for revealing the larger mixed-list benefit observed in our experiment. In particular, this manipulation appears to have resulted in a lower FA rate in the mixed-aloud condition than is obtained using the standard mixed-list versus pure-list design.

We contend that the FA rate in the mixed-aloud condition of Experiment 3 may been particularly low because participants in this condition were aware that only aloud items or new items

[14] As in the previous experiments, Macmillan and Creelman's (2004) 1/2N correction was used to adjust ceiling hit rates and floor FA rates. This correction was applied to eight hit rates and 14 FA rates in the mixed-aloud condition; zero hit rates and three FA rates in the mixed-silent condition; six hit rates and seven FA rates in the pure-aloud condition; and two hit rates and eight FA rates in the pure-silent condition.

would appear at test, which made the distinctiveness heuristic particularly diagnostic, helping participants to correctly reject new words on the basis that they lacked aloud information. This differs from a standard old/new recognition test (in which silent words are also present) where the distinctiveness heuristic is not as diagnostic and therefore may not be relied upon as heavily by participants. But given that participants also use the distinctiveness heuristic following a pure-aloud list to reduce FAs (Dodson & Schacter, 2001), why did we find a lower FA rate in the mixed-aloud condition than in the pure-aloud condition? One possibility is that the distinctiveness of aloud information was made salient to participants at study, which may have prompted them to use an aloudness distinctiveness heuristic at test. Another possibility is that the instructions in the mixed-aloud condition, informing participants that there would be no silent items at test, may also have made more evident to participants the utility of a distinctiveness heuristic, prompting them to rely on it more than did participants in the pure-aloud condition.

General Discussion

We conducted three experiments that examined the costs and benefits that can be revealed by comparing mixed-list versus pure-list experimental designs using the production effect as a "case study." In particular, we did this in the context of producing aloud versus reading silently at encoding and their effects on a subsequent recognition test. The aloud and silent means for the pure-list conditions served as baselines against which we measured the benefit of reading words aloud—and the cost of reading words silently—in a mixed list (see also Bodner et al., 2014).

Across these three experiments, we found consistent evidence of both mixed-list and pure-list production effects. In each of these experiments, the effect size of the mixed-list production effect was larger than the effect size of the pure-list production effect, in accord with a distinctiveness account and with published meta-analyses (Fawcett, 2013; Bodner et al., 2014). Importantly, we found evidence in these three experiments that this larger mixed-list production effect was partially due to a benefit of reading aloud in a mixed list versus a pure list, a result that has not been previously reported.

In our Experiment 1, participants studied three types of word lists in a random order: pure aloud, pure silent, and mixed. After each list, they were given an old/new recognition test. In terms of hit rates, we found a memorial benefit of reading aloud in a mixed list versus a pure list, as well as a cost to reading silently in a mixed list versus a pure list. A drawback of this design—the design routinely used in this realm—was that it was not possible to make corresponding mixed versus pure list comparisons for FA rates because only a single, combined FA rate could be obtained for the mixed list. Moreover, this combined mixed FA rate prevented us from obtaining accurate d' calculations in the mixed-aloud and mixed-silent conditions.

In Experiments 2 and 3, we addressed the FA issue by obtaining separate FA rates in the mixed-aloud and mixed-silent conditions. These separate mixed-list FA rates were obtained by making two different modifications to the mixed versus pure list experimental design. In Experiment 2, participants were instructed to make modality attributions at test. In Experiment 3, participants who studied a mixed list received a "pure" recognition test in which they were tested only on aloud studied words or only on silent studied words. The separate mixed FA rates obtained in these experiments allowed us to calculate independent d' values for their respective mixed-aloud and mixed-silent conditions, which in turn allowed us to assess costs and benefits in d'. Our patterns of results illustrate some of the problems inherent in the traditional approach to design comparisons, problems that go well beyond the production effect used here as a kind of "case study."

Importantly with respect to the production effect, we found a mixed-list d' benefit of reading aloud in both Experiments 2 and 3. Unlike our Experiment 1, in which a mixed-aloud benefit in hits was mitigated by the use of a combined mixed FA rate, the results of Experiments 2 and 3 showed a mixed-list benefit in hits that was accompanied by a corresponding benefit in FAs: Participants were less likely to misclassify new words as aloud following a mixed list than following a pure list. This combination of higher hits and lower FAs conforms to the well-known "mirror effect" pattern in recognition (see Glanzer & Adams, 1985; Murdock, 2003).

Like Bodner et al. (2014), we also found evidence of a cost to reading silently in a mixed list. In all three of our experiments, hit rates were lower for silent items that were studied in a mixed list versus a pure list. In terms of FA rates, however, we did not find reliable evidence of a mixed-list cost of reading silently. When we dissociated mixed-list FA rates (Experiment 2 and 3), FA rates in the mixed-silent condition were not consistently higher than they were in the pure-silent condition.[15] Thus, in terms of overall memory discrimination (d'), dissociating mixed-list FAs appears to have mainly had the effect of amplifying the mixed-list benefit, without having much of an effect on the mixed-list cost. This is important to be aware of when the purpose of a study—any study, not just one involving the production effect—is to examine costs and benefits using design comparisons.

Distinctiveness and the Benefit of Production

The present research demonstrated that the memorial advantage of mixed-list production applies not only to hit rates, but also to FA rates. In Experiments 2 and 3, participants who studied a mixed list were less likely to incorrectly judge new words to have been studied aloud versus silently. Moreover, we found lower FA rates in the mixed-aloud condition than in the pure-aloud condition. This mixed-list benefit for FAs mirrored the mixed-list benefit for hits. We submit that these mixed-list benefits may have arisen because the distinctiveness of aloud information was made salient at study, making it more accessible to participants at test than it was following a pure-aloud list.

To be clear, the results of Experiments 2 and 3 do not provide conclusive evidence that participants were more likely to FA silent than aloud to new words following a standard old/new recognition test, in which only a combined mixed FA rate can be obtained (e.g., the recognition test used in Experiment 1 here or in Experiment 1 of Bodner et al., 2014). It is impossible to know what participants were thinking when they misclassified a new word as old—they might have thought that they had studied the word aloud, or silently, or just that it felt familiar.

[15] An analysis of the first-block data of Experiment 2 showed some evidence of this trend but that was the only time that we saw it.

Although Experiments 2 and 3 may highlight a general tendency of individuals to be less likely to mistake new words as aloud than silent following a mixed list, it is also possible that this result was specific to the types of recognition tests used here. Perhaps the format of these recognition tests encouraged participants to use a speech distinctiveness heuristic. Ordinarily, the use of a speech distinctiveness heuristic is somewhat ineffective following a mixed study list because the absence of aloud information does not guarantee that a word is new (it could have been studied silently). Participants may still persist, however, in using this heuristic to some extent following a mixed list (cf. Dodson & Schacter, 2001, Experiment 2), which may result in them classifying silent words as new, consistent with the mixed-list cost observed here, and in Bodner and colleagues' (2014) research.

When participants were asked to indicate the modality in which they had studied a word (Experiment 2), or were tested only on aloud words following a mixed study list (Experiment 3), the distinctiveness heuristic became a more effective tool for rejecting lures, as the absence of an aloud record became stronger evidence that the word was not studied. In Experiment 2, the absence of an aloud record following a mixed study list helped participants narrow down their response options to silent or new. Additional information (perhaps the evaluated strength of the item; see Bodner & Taikh, 2012) would have then been required to differentiate between silent and new words. In Experiment 3, the distinctiveness heuristic was even more effective because on the aloud-only test the absence of aloud information constituted strong evidence that the word was new, given that there were no silent items on that test. Moreover, the fact that we informed participants in the mixed-aloud test condition that there would be no silent items at test may have made the utility of the aloudness distinctiveness heuristic especially salient.

Overall, participants may have been more inclined to use the distinctiveness heuristic in Experiments 2 and 3 than they ordinarily would be on a standard old/new recognition test (e.g., Experiment 1) because the use of this tool was less costly (i.e., it ought not to have increased the likelihood of them failing to correctly recognize silently studied words). This increased use of the distinctiveness heuristic may have augmented the benefit of reading aloud in a mixed list.

Interpreting the Cost of Mixed-List Production

As argued above, the use of a distinctiveness heuristic following a mixed study list may ordinarily impose a cost on silent items: Participants are biased to dismiss them as being new because they lack a distinct aloud record at test. In this sense, the distinctiveness heuristic may be a "double-edged sword" following a mixed list—enhancing recognition for aloud items while simultaneously impairing recognition for silent items (comparable results were obtained by Huff et al., 2015, using a DRM paradigm). This may account for the mixed-list cost to the correct recognition of silent items in Experiment 1.

In Experiment 2, this cost was reduced; indeed, in both the full data and first-block data, the cost was nonsignificant in terms of hits. This may have been because there was a 'silent' response option on the recognition test, which could have discouraged participants from dismissing as being new words that lacked an aloud record. In Experiment 3, however, participants were given a pure test following the mixed study list. Participants in the mixed-silent condition knew that *only* silent and new items would appear on the recognition test. Therefore, participants would have no reason to employ an aloudness distinctiveness heuristic on this pure test, and hence their discrimination of silent items, in theory, should not have been compromised. Yet we found a statistically significant cost in terms of hits in Experiment 3.

If the distinctiveness heuristic cannot explain the cost in Experiment 3, then what can? Given that the tests for both mixed-silent and pure-silent items were equivalent, it appears that the cost to mixed-silent items was imposed at study. Lazy reading (cf. Begg & Snider, 1987) would seem to be a prime candidate for explaining this cost: Relative to the distinctively processed aloud words, participants might tend to shallowly process the silent words on a mixed list, judging the words they read aloud to be more important. This account does not accord well, though, with a finding recently reported by Forrin, Jonker, and MacLeod (2014)—that there is a robust production effect even when participants elaboratively process both aloud and silent words, whether through generation or imagery—ensuring that the silent words are not lazily read (see also Experiments 7 and 8 in MacLeod et al., 2010). Bodner et al. (2014), however, found that blocking aloud and silent words at study eliminated the cost of mixed-list production, a result which they noted was consistent with a lazy reading account. Overall, then, the evidence regarding the lazy reading account has been mixed. Further research is needed to investigate whether lazy reading plays a role in the cost to shallowly processed items in a mixed list.

Evaluating the Viability of Mixed-List Versus Pure-List Designs for Assessing Costs and Benefits

We have argued here that the standard mixed-list versus pure-list design is not ideal for assessing costs versus benefits in recognition research because separate mixed-list FA rates cannot be obtained. Experiments 2 and 3 made modifications that dissociated mixed-list FA rates. Are these designs therefore better suited to assessing costs and benefits? Not necessarily. Obtaining separate FAs may represent a step in the right direction, but these experiments still had issues worth noting.

In Experiment 2, in which participants made modality attributions, the mixed-list recognition test (in which participants had three response options) was arguably more difficult than the pure-list recognition tests (in which participants had two response options), resulting in lower mixed-list hit rates. Our solution to this predicament was to score hit rate data as "old" or "new," essentially doing away with the modality data that this experiment afforded us when comparing hit rates. This analytic approach yielded a mixed-list benefit in terms of hit rates, consistent with the results of Experiments 1 and 3. When hits were more conservatively scored as requiring correct modality judgments, mixed-aloud and mixed-silent hit rates were notably lower, shifting the results from mixed-list benefit to mixed-list cost.

Thus, a downside of this design is that, although the modality attributions allowed for a direct comparison of mixed-list and pure-list FAs, these same attributions made it difficult to directly compare mixed-list and pure-list hit rates. The analyses that we opted for (converting aloud and silent responses into old responses) may have biased our results in the direction of a mixed-

list benefit of reading aloud because we were scoring silent recognition responses to aloud words as hits.

In Experiment 3, we addressed this issue by testing participants who had studied a mixed list using a "pure" recognition test, which included only aloud or only silent words from study (along with new words). This design allowed separate hit rates and separate FA rates to be obtained following a mixed study list, which could then be compared to the hit rates and FA rates from the pure lists. In this manner, study list type (mixed vs. pure) was the only factor manipulated; the recognition test was identical in each condition. Although there may be no "gold-standard" of assessing costs and benefits in recognition, a pure-test design may constitute the most accurate—and hence best available—approach.

The Advantages of Dissociating Mixed-List FA Rates

We have argued that dissociating the mixed-list FA rates allows for a more accurate assessment of costs and benefits because this approach permits comparison of mixed-list and pure-list FA rates. Even when such comparisons are not of primary interest, however, we still advocate the use of recognition tests that allow for the dissociation of mixed-list FAs, irrespective of the encoding technique that is being investigated. Obtaining separate mixed-list FAs permits the calculation of memory discrimination for each type of item that was studied. Moreover, our result that individuals were less likely to FA aloud than silent to new words suggests that individuals may use a distinctiveness heuristic on recognition tests on which they can indicate study modality (Experiment 2), or on which only one type of studied item appears (Experiment 3). Indeed, individuals may be particularly inclined to employ a distinctiveness heuristic following a mixed list in which a certain item type (e.g., aloud words) receives distinctive processing.

An advantage specific to having participants make modality attributions at test is that the modality data may prove informative with respect to assessing theoretical accounts. In Experiment 2, for example, we found that participants were more likely to misclassify words that they had studied aloud as silent as opposed to new, a result that was not consistent with a strong distinctiveness account: If participants were only basing their modality decision on the presence/absence of a distinct aloud record, then they should have been equally likely to misclassify aloud words as new or as silent. Instead, this modality result aligned well with Bodner and Taikh's (2012) evaluative strength account. Participants may use strength/familiarity judgments (perhaps in addition to distinctiveness judgments) to guide their recognition decisions.

Summary and Conclusion

Across three experiments, we found that there is a memorial benefit to reading aloud in a mixed list versus a pure list.[16] In each experiment, this benefit was evident in terms of hit rates. In Experiments 2 and 3, in which we used recognition tests that yielded separate mixed-list FA rates, we found consistent evidence that this mixed-list benefit also extended to FAs (participants were less likely to miscategorize new words as having been studied aloud following a mixed vs. pure study list). As well, we observed a mixed-list benefit in memory discrimination (d') in Experiments 2 and 3. The usual benefit in hits was accompanied by the benefit in FAs in these two experiments, resulting in an overall mixed-list benefit in d'.

We also found evidence of a mixed-list cost of reading silently (see also Bodner et al., 2014), although this evidence was less consistent across the three experiments than was the evidence for a benefit. A cost was mainly evident in our hit rate data. There was limited evidence in Experiments 2 and 3 that there was a mixed-list cost to reading silently in terms of higher FA rates. Contrary to Bodner and colleagues' meta-analysis, we also found limited evidence of a mixed-list d' cost (we found a significant d' cost only once—in the first-block analysis of Experiment 2, shown in Appendix B). Our present research suggests that the cost of reading silently in a mixed list is not increased by dissociating mixed-list FA rates.

We argue that the mixed-list benefit of reading aloud arises due to distinctiveness: During study, aloud words stand out as distinct on a mixed list, which may increase the likelihood that participants adopt a distinctiveness heuristic at test following a mixed list as opposed to a pure list. Given that recognition is influenced by contextual factors (e.g., Bodner & Lindsay, 2003), another possibility is that words that are read aloud benefit from "stronger" encoding in a mixed-list (against a context of silent items) than in a pure list. Participants may then evaluate the strength (Bodner & Taikh, 2012) of items at test to help guide their recognition decisions. The modality data in our Experiment 2 provided novel support for the evaluated strength account.

The mechanism underlying the cost to reading silently in a mixed list is not yet clear. The modality attribution results of Experiment 2 suggest that participants have a propensity to judge silent words as new, perhaps because they use the distinctiveness heuristic at test and, finding no aloud record, take this absence as evidence that a word was not studied. Another candidate mechanism is lazy reading (Begg & Snider, 1987), but the evidence for this in the literature is mixed.

From a practical standpoint, these results support the conclusion that reading aloud is a worthwhile study strategy (see Ozubko, Hourihan, & MacLeod, 2012), especially when that reading aloud occurs against a context of silent reading. As is frequently reported by people who have extensive experience with studying, saying the important parts aloud really seems to help in remembering them, both in remembering that something was studied and in remembering that something was not studied.

In sum, the present results demonstrate that there is a robust mixed-list benefit of reading aloud—both in terms of hits and FAs (when the recognition test allows separate FAs to be obtained for the mixed-list conditions). Importantly, these results have theoretical implications that go beyond the production effect: They demonstrate that an assessment of the costs and benefits of a given encoding technique may depend heavily on the type of recognition test used. Recognition tests that dissociate mixed-list FA rates appear to yield a more pronounced benefit, perhaps because they encourage participants to employ a distinctiveness heuristic at test—a heuristic that is ordinarily less effective in a mixed design (see Dodson & Schacter, 2001, Experiment 2). As always in

[16] Although this article has focused on recognition, other research has examined the costs and benefits of mixed-list production in recall (see Forrin & MacLeod, 2015; Jones & Pyc, 2014; Jonker, Levene, and MacLeod, 2014). Comparison of these recognition vs. recall results falls outside the scope of this article, but the interested reader can see Forrin and MacLeod (2015) for a discussion.

experimental psychology, finding the best way to measure a phenomenon requires considerable thought, and subtle changes in procedure can have important implications both for measurement and for theory.

References

Begg, I., & Roe, H. (1988). On the inhibition of reading by generating. *Canadian Journal of Psychology/Revue canadienne de psychologie, 42,* 325–336. http://dx.doi.org/10.1037/h0084191

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

Begg, I., Snider, A., Foley, F., & Goddard, R. (1989). The generation effect is no artifact: Generating makes words distinctive. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 15,* 977–989. http://dx.doi.org/10.1037/0278-7393.15.5.977

Bertsch, S., Pesta, B. J., Wiscott, R., & McDaniel, M. A. (2007). The generation effect: A meta-analytic review. *Memory & Cognition, 35,* 201–210. http://dx.doi.org/10.3758/BF03193441

Bodner, G. E., & Lindsay, D. S. (2003). Remembering and knowing in context. *Journal of Memory and Language, 48,* 563–580. http://dx.doi.org/10.1016/S0749-596X(02)00502-8

Bodner, G. E., & Taikh, A. (2012). Reassessing the basis of the production effect in memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 38,* 1711–1719. http://dx.doi.org/10.1037/a0028466

Bodner, G. E., Taikh, A., & Fawcett, J. M. (2014). Assessing the costs and benefits of production in recognition. *Psychonomic Bulletin & Review, 21,* 149–154. http://dx.doi.org/10.3758/s13423-013-0485-1

Conway, M. A., & Gathercole, S. E. (1987). Modality and long-term memory. *Journal of Memory and Language, 26,* 341–361. http://dx.doi.org/10.1016/0749-596X(87)90118-5

Dodson, C. S., & Schacter, D. L. (2001). "If I had said it I would have remembered it": Reducing false memories with a distinctiveness heuristic. *Psychonomic Bulletin & Review, 8,* 155–161. http://dx.doi.org/10.3758/BF03196152

Engelkamp, J., & Zimmer, H. D. (1997). Sensory factors in memory for subject-performed tasks. *Acta Psychologica, 96,* 43–60. http://dx.doi.org/10.1016/S0001-6918(97)00005-X

Erdfelder, E., Faul, F., & Buchner, A. (1996). GPOWER: A general power analysis program. *Behavior Research Methods, Instruments, & Computers, 28,* 1–11.

Fawcett, J. M. (2013). The production effect benefits performance in between-subject designs: A meta-analysis. *Acta Psychologica, 142,* 1–5. http://dx.doi.org/10.1016/j.actpsy.2012.10.001

Forrin, N. D., Jonker, T. R., & MacLeod, C. M. (2014). Production improves memory equivalently following elaborative vs non-elaborative processing. *Memory, 22,* 470–480. http://dx.doi.org/10.1080/09658211.2013.798417

Forrin, N. D., & MacLeod, C. M. (2015). *Order information guides recall of silent items studied in long lists*. Manuscript in preparation.

Gathercole, S. E., & Conway, M. A. (1988). Exploring long-term modality effects: Vocalization leads to best retention. *Memory & Cognition, 16,* 110–119. http://dx.doi.org/10.3758/BF03213478

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

Hopkins, R. H., & Edwards, R. E. (1972). Pronunciation effects in recognition memory. *Journal of Verbal Learning & Verbal Behavior, 11,* 534–537. http://dx.doi.org/10.1016/S0022-5371(72)80036-7

Huff, M. J., Bodner, G. E., & Fawcett, J. M. (2015). Effects of distinctive encoding on correct and false memory: A meta-analytic review of costs and benefits and their origins in the DRM paradigm. *Psychonomic Bulletin & Review, 22,* 349–365. http://dx.doi.org/10.3758/s13423-014-0648-8

Hunt, R. R. (2006). The concept of distinctiveness in memory research. In R. R. Hunt & J. B. Worthern (Eds.), *Distinctiveness and memory* (pp. 1–25). New York, NY: Oxford University Press. http://dx.doi.org/10.1093/acprof:oso/9780195169669.003.0001

Hunt, R. R. (2013). Precision in memory through distinctive processing. *Current Directions in Psychological Science, 22,* 10–15. http://dx.doi.org/10.1177/0963721412463228

Hunt, R. R., & Worthen, J. B. (2006). *Distinctiveness and memory*. New York: Oxford University Press. http://dx.doi.org/10.1093/acprof:oso/9780195169669.001.0001

Jones, A. C., & Pyc, M. A. (2014). The production effect: Costs and benefits in free recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 40,* 300–305. http://dx.doi.org/10.1037/a0033337

Jonides, J., & Mack, R. (1984). On the cost and benefit of cost and benefit. *Psychological Bulletin, 96,* 29–44. http://dx.doi.org/10.1037/0033-2909.96.1.29

Jonker, T. R., Levene, M., & Macleod, C. M. (2014). Testing the item-order account of design effects using the production effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 40,* 441–448. http://dx.doi.org/10.1037/a0034977

MacLeod, C. M., Gopie, N., Hourihan, K. L., Neary, K. R., & Ozubko, J. D. (2010). The production effect: Delineation of a phenomenon. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36,* 671–685. http://dx.doi.org/10.1037/a0018785

Macmillan, N. A., & Creelman, C. D. (2004). *Detection theory: A user's guide* (2nd ed.). Mahwah, NJ: Erlbaum.

McDaniel, M. A., & Bugg, J. M. (2008). Instability in memory phenomena: A common puzzle and a unifying explanation. *Psychonomic Bulletin & Review, 15,* 237–255. http://dx.doi.org/10.3758/PBR.15.2.237

McDaniel, M. A., & Einstein, G. O. (1986). Bizarre imagery as an effective memory aid: The importance of distinctiveness. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 12,* 54–65. http://dx.doi.org/10.1037/0278-7393.12.1.54

Murdock, B. (2003). The mirror effect and the spacing effect. *Psychonomic Bulletin & Review, 10,* 570–588. http://dx.doi.org/10.3758/BF03196518

Ozubko, J. D., Gopie, N., & MacLeod, C. M. (2012). Production benefits both recollection and familiarity. *Memory & Cognition, 40,* 326–338. http://dx.doi.org/10.3758/s13421-011-0165-1

Ozubko, J. D., Hourihan, K. L., & MacLeod, C. M. (2012). Production benefits learning: The production effect endures and improves memory for text. *Memory, 20,* 717–727. http://dx.doi.org/10.1080/09658211.2012.699070

Ozubko, J. D., Major, J., & MacLeod, C. M. (2014). Remembered study mode: Support for the distinctiveness account of the production effect. *Memory, 22,* 509–524. http://dx.doi.org/10.1080/09658211.2013.800554

Slamecka, N. J., & Katsaiti, L. T. (1987). The generation effect as an artifact of selective displaced rehearsal. *Journal of Memory and Language, 26,* 589–607. http://dx.doi.org/10.1016/0749-596X(87)90104-5

Thorndike, E. L., & Lorge, I. (1944). *The teacher's word book of 30,000 words*. New York, NY: Columbia University, Teachers College.

Wickelgren, W. A. (1969). Associative strength theory of recognition memory for pitch. *Journal of Mathematical Psychology, 6,* 13–61. http://dx.doi.org/10.1016/0022-2496(69)90028-5

(*Appendices follow*)

Appendix A

Additional Analyses for Experiment 1

Order Effects

First, we tested whether participants' memory discrimination (d') worsened across study-test blocks, as they contended with an increased amount of proactive interference from previous blocks. (For the mixed block, we averaged participants' mixed-aloud and mixed-silent d' values.) Not surprisingly, mean d' was significantly higher in the first block ($M = 2.11$, $SE = 0.06$) than in the second block ($M = 1.91$, $SE = 0.06$), $t(133) = 2.73$, $p = .007$, $d = 0.29$. Mean d' was also higher in the second block than in the third block ($M = 1.79$, $SE = 0.07$), although this trend was not significant, $t(133) = 1.56$, $p = .12$, $d = 0.16$. Hits rates showed the same pattern: Hits rates were lower in the second block ($M = .74$, $SE = 0.01$) than in the first block ($M = .77$, $SE = 0.01$), $t(133) = 2.24$, $p = .03$, $d = 0.20$, and in the third block ($M = .71$, $SE = 0.01$) than in the second block, $t(133) = 2.48$, $p = .01$, $d = 0.22$.

Consistent with the above pattern of worsening memory discrimination across blocks, we found that the pure-list production effect was significantly larger in block orders in which the pure-aloud block was presented before the pure-silent block ($M = 0.77$, $SE = 0.12$), compared to orders in which the pure-aloud block was presented after the pure-silent block ($M = 0.30$, $SE = 0.10$), $t(132) = 3.14$, $p = .002$, $d = 0.55$. Along the same lines, the mixed-list benefit was also larger when the mixed-block preceded ($M = 0.36$, $SE = 0.09$), rather than followed ($M = -0.21$, $SE = 0.09$) the pure-aloud block, $t(132) = 4.38$, $p < .001$, $d = 0.76$. And the mixed-list cost was significantly larger when the pure-silent block preceded ($M = -0.22$, $SE = 0.09$), rather than followed ($M = 0.12$, $SE = 0.11$) the mixed block, $t(132) = 2.37$, $p = .02$, $d = 0.41$. Overall, the size of the mixed-list production effect did not significantly differ across the three blocks, $F(2, 131) = 2.03$, $MSE = 0.28$, $p = .14$, $\eta^2 = 0.03$, which was expected given that the mixed-list production effect was a within-block effect—for which both conditions would have been equally affected by proactive interference or fatigue—in contrast to the other three effects, which were *between-blocks.* All of these effects were identical when we analyzed hit rates rather than d'.

The above order effects are consistent with memory performance decreasing across the three study-test blocks. Our counterbalancing of the block order should have effectively neutralized the impact of order effects, which likely explains why our results were consistent for the full data analyses (reported in the body of the article) and the first-block data analyses (reported below). These analyses were conducted on all 134 participants' first block data (mixed: $n = 44$; pure aloud: $n = 45$; pure silent: $n = 45$).

Hit Rates

There was a significant mixed-list production effect in terms of hit rates: Participants who studied a mixed list correctly recognized a greater proportion of aloud words than silent words, $t(43) = 9.37$, $p < .001$, $d = 1.39$. Participants who studied a pure-aloud list also had higher hit rates compared to those who studied a pure-silent list, $t(88) = 3.69$, $p < .001$, $d = 0.79$. There was a significant benefit to reading words aloud in a mixed-list compared to a pure-aloud list, $t(87) = 2.20$, $p < .05$, $d = 0.47$, and also a significant cost to reading silently in a mixed list relative to a pure-silent list, $t(87) = 2.01$, $p < .05$, $d = 0.43$.

FA Rates

Only an overall FA rate could be obtained for participants who studied a mixed list ($M = 0.12$, $SE = 0.01$), which was used as an estimate of both the mixed-aloud and mixed-silent FA rates. In terms of pure-list FA rates, participants had significantly lower FA rates for pure-aloud words relative to pure-silent words, $t(76.58) = 2.32$, $p < .05$, $d = 0.49$. In terms of a mixed-list benefit, the overall mixed-list FA rate was not reliably different from the pure-aloud FA rate, $t(87) = 0.52$, *ns*. In terms of a mixed-list cost, the overall mixed-list FA rate was marginally lower than the pure-silent FA rate, $t(87) = -1.76$, $p = .08$, $d = 0.38$, a modest-sized effect.

Memory Discrimination (d')

For computing d', Macmillan and Creelman's (2004) 1/2N correction was used to adjust ceiling hit rates and floor FA rates. This correction was applied to 7 aloud hit rates, 0 silent hit rates, and 6 FA rates in the mixed condition; 1 hit rate and 3 FA rates in the pure-aloud condition; and 0 hit rates and 2 FA rates in the pure-silent condition.

Only a single FA rate could be obtained for participants who studied a mixed list, which therefore had to be used to provide an estimate for calculating memory discrimination (d') in both the mixed-aloud and mixed-silent conditions. Using this estimate, participants had superior memory discrimination for words studied aloud compared to words studied silently in a mixed list, $t(43) = 10.12$, $p < .001$, $d = 1.13$. Participants also showed superior memory discrimination for pure-aloud words compared to pure-silent words, $t(133) = 6.86$, $p < .001$, $d = 0.74$. There was not a significant benefit associated with reading aloud in a mixed list relative to a pure-aloud list, $t(87) = 1.48$, *ns*, nor was there a significant cost, $t(80.23) = 0.12$, *ns*. These benefit and cost d' results should be interpreted cautiously because the overall mixed-list FA rate was used to calculate both the mixed aloud d' value and the mixed-silent d' value.

(*Appendices continue*)

Appendix B

Additional Analyses for Experiment 2

Order Effects

Consistent with the previous experiment, participants' memory performance also decreased across the three qblocks of Experiment 2. Participants' memory discrimination was significantly higher in the first block ($M = 2.33$, $SE = 0.06$) than in the second block ($M = 2.04$, $SE = 0.07$), $t(133) = 3.68$, $p < .001$, $d = 0.39$; the second block did not differ from the third block ($M = 2.02$, $SE = 0.07$), $t(133) = 0.19$, *ns*. An analysis of hit rates revealed an identical pattern, as was the case for the results described below.

Once again, we found order effects consistent with the above pattern of worsening memory discrimination across blocks. The pure-list production effect was significantly larger in block orders in which the pure-aloud block preceded the pure-silent block ($M = 0.85$, $SE = 0.11$), compared to when it followed the pure-silent block ($M = 0.43$, $SE = 0.11$), $t(132) = 2.67$, $p = .01$, $d = 0.46$. Similarly, the mixed-list benefit was larger when the mixed-list block preceded ($M = 0.86$, $SE = 0.09$), rather than followed ($M = 0.39$, $SE = 0.08$), the pure-aloud block, $t(132) = 3.82$, $p < .001$, $d = 0.66$. And the mixed-list cost was significantly larger when the pure-silent block preceded ($M = -0.31$, $SE = 0.09$), rather than followed ($M = 0.20$, $SE = 0.13$), the mixed-list block, $t(117.66) = 3.11$, $p = .002$, $d = 0.54$. Overall, the size of the mixed-list production effect did not significantly differ across the three blocks, $F(2, 131) = 0.98$, $MSE = 0.96$, $p = .38$, $\eta^2 = 0.01$, consistent with the fact that the mixed-list production effect was a within-block effect that would not have been affected by memory performance decreasing across blocks.

As was the case with Experiment 1, the order effects in Experiment 2 were consistent with memory performance decreasing across the three study-test blocks. Thus, the block counterbalancing should have ensured that these order effects did not bias our results in the direction of either costs or benefits. We analyzed the data from each participant's first experimental block (mixed: $n = 45$; pure aloud: $n = 44$; pure silent: $n = 45$); these analyses are reported below.

Hit Rates

In terms of mixed-list hit rates, a greater proportion of aloud words than silent words were recognized as having been studied, $t(44) = 7.21$, $p < .001$, $d = 1.23$. There was also a significant pure-list production effect, $t(87) = 2.87$, $p < .01$, $d = 0.61$. Compared to the pure-aloud condition, the mixed-aloud condition had a significantly higher proportion of hits, $t(87) = 2.00$, $p < .05$, $d = 0.43$, reflecting a benefit of mixed-list production. Hit rates tended to be lower for mixed-silent items than for pure-silent items, although this cost was not statistically significant, $t(88) = 1.47$, $p = .14$, $d = 0.31$.

FA Rates

In terms of mixed-list FA rates, a greater proportion of aloud words than silent words were recognized as having been studied, $t(44) = 7.21$, $p < .001$, $d = 1.23$. Although there was a lower proportion of FAs for pure aloud items (.100) than for pure silent items (.129), this difference was not statistically significant, $t(87) = 1.28$, $p = .21$, $d = 0.27$. (This analysis diverged from that on the full data, reported in the Results section of Experiment 2, in which there was a reliably lower proportion of FAs for the pure-aloud words.) Compared to the pure-aloud condition, the mixed-aloud condition had a significantly lower proportion of FAs, $t(67.42) = 4.30$, $p < .001$, $d = 0.91$, reflecting a mixed-list benefit. FAs tended to be higher for mixed-silent items than for pure-silent items, consistent with a cost, but this difference also was not reliable, $t(88) = 1.47$, $p = .14$, $d = 0.31$.

Memory Discrimination (d')

For computing d', ceiling hit rates and floor FA rates were adjusted using Macmillan and Creelman's (2004) 1/2N correction. This correction was applied to sixhit rates and 24 FA rates in the mixed-aloud condition; zero hit rates and three FA rates in the mixed-silent condition; zero hit rates and six FA rates in the pure-aloud condition; and zero hit rates and eight FA rates in the pure-silent condition. This large number of FA adjustments arguably yielded a conservative calculation of d' in the mixed-aloud condition relative to the other conditions of this experiment, which did not have as large of a disparity between ceiling hit rates and floor FA rates. Nevertheless, a robust d' benefit was observed.

Overall, memory discrimination (d') was superior for mixed-aloud versus mixed-silent items, $t(44) = 9.42$, $p < .001$, $d = 2.15$, a robust mixed-list production effect. Memory discrimination was also superior for pure-aloud versus pure-silent items, $t(87) = 2.41$, $p < .05$, $d = 0.52$. There was a robust benefit in d' for reading aloud in a mixed list compared to a pure list, $t(87) = 5.15$, $p < .001$, $d = 1.09$. There was a significant d' cost of reading silently in a mixed list compared to a pure list, $t(88) = 2.05$, $p < .05$, $d = 0.43$, although the cost effect size was modest relative to that of the large d' benefit. The significant d' cost was the result of trends for both hits and FAs being in the direction of a mixed-list cost.

Of note, this cost was only evident in the first-block data, and not in the full data reported in the Results section of Experiment 2. This is a somewhat puzzling difference given the otherwise high degree of concordance between our full analyses and first-block analyses. On the one hand, our analyses of the block order effects (reported above) did not provide any evidence that participants' results were biased in any way that would have minimized the mixed-list cost of reading silently. On the other hand, the mixed-list cost revealed by the first-block data is consistent with the modest d' cost reported in Bodner and colleagues' (2014) meta-analysis. Thus, there may be a modest mixed-list d' cost to reading silently, although it was absent in Experiment 1, and not consistently present in Experiment 2.

(*Appendices continue*)

Modality Attributions (First-Block Data)

We reanalyzed costs and benefits using the more stringent criterion of only scoring correct modality attributions as hits. Here we report the results of the first-block data, which were identical to the results of the full data. Not surprisingly, using this stringent scoring criterion considerably reduced the hit rates for the mixed-aloud and mixed-silent conditions. These lower hit rates resulted in an even more pronounced mixed-list cost to reading silently. The mixed-silent hit rate ($M = 0.57$, $SE = 0.02$) was significantly lower than the pure-silent hit rate ($M = 0.75$, $SE = 0.02$), $t(88) = 6.40$, $p < .001$, $d = 1.36$, and the mixed-silent d' ($M = 1.30$, $SE = 0.11$) was significantly lower than the pure-silent d' ($M = 2.06$, $SE = 0.13$), $t(88) = 4.59$, $p < .001$, $d = 0.98$.

These analyses yielded a mixed-list cost for reading aloud, in contrast to the benefit that was evident when 'aloud' and 'silent' responses were collapsed into 'old' responses. When hits were scored stringently, the mixed-aloud hit rate ($M = 0.64$, $SE = 0.03$) was significantly lower than the pure-aloud hit rate ($M = 0.82$, $SE = 0.01$), $t(87) = 5.81$, $p < .001$, $d = 1.25$, and the mixed-aloud d' ($M = 2.37$, $SE = 0.11$) was lower than the pure-aloud d' ($M = 2.45$, $SE = 0.10$), although this difference was not statistically significant, $t(87) = 0.58$. Notably, there was still a significant mixed-list production effect when hits were coded as correct modality attributions—in terms of both hits, $t(44) = 2.05$, $p < .05$, $d = 0.42$, and d', $t(44) = 7.11$, $p < .001$, $d = 1.48$.

In terms of the mixed-list modality data, the pattern of results for the first-block data were identical to those of the full data. Of note, participants were less likely to FA aloud than silent to new words, $t(44) = 7.21$, $p < .001$, $d = 1.23$. Participants were also less likely to miscategorize silent words as aloud than as new, $t(44) = 5.07$, $p < .001$, $d = 1.18$. These two results suggest that participants are less likely to confuse aloud words with new words than they are to confuse silent words with new words, which is consistent with participants using an aloudness distinctiveness heuristic at test. However, participants were also more likely to miscategorize aloud words as silent than as new, $t(44) = 3.28$, $p = .002$, $d = 0.63$, a result inconsistent with participants relying strictly on a distinctiveness heuristic.

Comparing the Mixed-List Benefits in Experiments 1 and 2 (First-Block Data)

To compare the mixed-list benefit of reading aloud in Experiments 1 and 2, we conducted a two-way mixed-model ANOVA with study list type (mixed-aloud vs. pure-aloud) and experiment (Experiment 1 first block vs. Experiment 2 first block) both as between-subjects factors. For hit rates, the two-way ANOVA revealed a significant main effect of study list type, $F(1, 174) = 8.81$, $MSE = 0.01$, $p = .003$, $\eta^2 = 0.05$, reflecting an overall mixed-list benefit of reading aloud in terms of hits. The main effect of Experiment, however, was nonsignificant, as was the Experiment × Study List Type interaction (both Fs < 1), suggesting that the two experiments yielded comparable mixed-list benefits in terms of hits (when hits were scored leniently in Experiment 2).[17]

Next, we ran this ANOVA with FA rates as the dependent measure. The main effect of study list type was significant, $F(1, 174) = 6.16$, $MSE = 0.01$, $p = .01$, $\eta^2 = 0.03$, signifying overall lower FA rates in the mixed-aloud condition than in the pure-aloud condition. The main effect of experiment was also significant, $F(1, 174) = 10.58$, $MSE = 0.01$, $p < .001$, $\eta^2 = 0.08$, reflecting overall lower FA rates in the aloud conditions of Experiment 2 than Experiment 1. Of primary interest, the Experiment × Study List Type interaction was significant, $F(1, 174) = 10.58$, $MSE = 0.01$, $p = .001$, $\eta^2 = 0.06$. FA rates in the mixed-aloud condition were significantly lower in Experiment 2 than they were in Experiment 1, $t(67.92) = 5.41$, $p < .001$, $d = 1.15$, whereas FA rates in the pure-aloud condition did not differ significantly between experiments, $t(87) = 0.43$.

Last, we ran this ANOVA with d' as the dependent measure. An identical pattern of results was observed as reported above with FAs. The ANOVA yielded a robust main effect of study list type, $F(1, 174) = 21.62$, $MSE = 0.48$, $p < .001$, $\eta^2 = 0.11$, signifying an overall benefit of reading aloud in a mixed-list versus a pure-list. The main effect of experiment was also significant, $F(1, 174) = 12.14$, $MSE = 0.48$, $p = .001$, $\eta^2 = 0.07$, reflecting overall higher d' values in the aloud conditions of Experiment 2 than Experiment 1. More important, the Experiment × Study List Type interaction was significant, $F(1, 174) = 6.34$, $MSE = 0.48$, $p = .01$, $\eta^2 = 0.04$, indicating a larger mixed-list benefit of reading aloud in Experiment 2 than in Experiment 1. Average mixed-aloud d' was also higher in Experiment 2 than in Experiment 1, $t(87) = 4.19$, $p < .001$, $d = 0.90$, whereas average pure-aloud d' did not differ significantly between experiments, $t(87) = 0.69$.[18]

[17] When we stringently scored hit rates in Experiment 2, the two-way ANOVA revealed a significant main effect of study list type, $F(1, 174) = 10.78$, $MSE = 0.02$, $p = 0.001$, $\eta^2 = 0.06$, signifying that mixed-aloud hit rates were lower overall than pure-aloud hit rates (in contrast to the mixed-list benefit that arose when hits were leniently scored). The main effect of experiment was also significant, $F(1, 174) = 32.19$, $MSE = 0.02$, $p < 0.001$, $\eta^2 = 0.16$, indicating that aloud hit rates tended to be lower overall in Experiment 2 than in Experiment 1. The Experiment × Study List Type interaction was also robust, $F(1, 174) = 35.43$, $MSE = 0.02$, $p < 0.001$, $\eta^2 = 0.17$, reflecting the fact that there was a mixed-list benefit to reading aloud in Experiment 1, $t(87) = 2.20$, $p < 0.05$, $d = 0.47$, but a significant mixed-list cost to reading aloud in Experiment 2 (when hits were stringently scored), $t(87) = 5.81$, $p < 0.001$, $d = 1.25$. We contend that hit rates in the mixed-aloud condition of Experiment 2 were particularly low due to the added difficulty of making correct modality attributions, which involved three response choices instead of two.

[18] When stringently scored hit rates were used to derive the mixed-aloud d' values in Experiment 2, the ANOVA yielded non-significant main effects for study list type and for experiment (Fs < 1). The Experiment × Study List Type interaction was also nonsignificant, $F(1, 174) = 2.13$, $MSE = 0.50$, $p = 0.15$, $\eta^2 = 0.01$.

(*Appendices continue*)

Comparing the Mixed-List Costs in Experiments 1 and 2 (first-block Data)

The above results are consistent with our prediction that a d' benefit would emerge in Experiment 2 due to a decrease in the mixed-aloud FA rate relative to Experiment 1. But was the mixed-list cost of reading silently also larger in Experiment 2 than in Experiment 1? To address this question, we ran a series of two-way ANOVAs—parallel to those above—in which study list type (mixed-silent vs. pure-silent) and experiment (Experiment 1 first block vs. Experiment 2 first block) were between-subjects factors.

First, we ran this ANOVA with hits (leniently scored) as the dependent factor. The ANOVA yielded a significant main effect of Study List Type, $F(1, 175) = 6.21$, $MSE = 0.02$, $p = .01$, $\eta^2 = 0.03$, reflecting an overall mixed-list cost of reading silently. There was also a significant main effect of experiment, $F(1, 175) = 3.97$, $MSE = 0.02$, $p = .048$, $\eta^2 = 0.02$, indicating overall higher hit rates for silent items in Experiment 2 than in Experiment 1. The Experiment × Study List Type interaction, however, was not reliable ($F < 1$).[19]

In terms of fa rates, a two-way anova revealed nonsignificant main effects of both study list type and experiment (Fs < 1). The Experiment × Study List Type interaction was significant, however, $F(1, 175) = 5.11$, $MSE = 0.01$, $p = .03$, $\eta^2 = 0.03$. FA rates in the mixed-silent condition were significantly higher in Experiment 2 than in Experiment 1, $t(79.91) = 2.13$, $p = .04$, $d = 0.45$, whereas their pure-silent FA rates did not reliably differ between experiments, $t(88) = 1.10$, *ns*.

Last, we ran this ANOVA with d' as the dependent measure. This two-way ANOVA yielded a nonsignificant main effect of study list type, $F(1, 175) = 2.37$, $MSE = 0.47$, $p = .13$, $\eta^2 = 0.01$, showing that, across both experiments, there was no mixed-list cost to reading silently. The main effect of experiment was also nonsignificant, $F(1, 175) = 1.11$, $MSE = 0.47$, $p = .29$, $\eta^2 = 0.01$. Of main interest to us, the Experiment × Study List Type interaction was marginally significant, $F(1, 175) = 2.86$, $MSE = 0.01$, $p = .09$, $\eta^2 = 0.02$. As previously reported, the mixed-list cost of reading silently was statistically significant in the first-block data of Experiment 2, whereas this cost was nonsignificant in the first-block data of Experiment 1. Unexpectedly, however, this mixed-list cost did not arise in Experiment 2 because of a decrease in memory discrimination in the mixed-silent condition. Participants' memory discrimination in the mixed-silent condition was not reliably different between the two experiments ($t < 1$). Participants' memory discrimination in the pure-silent condition was higher in Experiment 2 than in Experiment 1, $t(88) = 1.91$, $p = .06$, $d = 0.41$; this trend was unexpected given that these conditions were essentially identical across the two experiments.[20]

Based on the above results, it is difficult to interpret the d' cost in the first-block data of Experiment 2. When mixed-list hit rates were scored leniently in Experiment 2, memory discrimination in the mixed-silent condition was not significantly worse in Experiment 2 than in Experiment 1, even though the mixed-silent FA rate was higher in Experiment 2 (this FA increase was offset by the fact that hit rate for mixed-silent items was also higher in Experiment 2). Unexpectedly, this cost seems to have arisen from an increase in the pure-silent d' in Experiment 2 relative to Experiment 1, despite the fact that the pure-silent conditions were essentially identical in the two experiments. Moreover, in the case of the full data (reported in the Results section of Experiment 2) the mean d' in the pure-silent condition did not significantly increase across experiments, and there was not a significant mixed-list cost in d'. Taken together, these results cast doubt on whether the design change implemented in Experiment 2 increased the mixed-list cost of reading silently.

The evidence was stronger that this design change did increase the mixed-list benefit of reading aloud. It appears that the design change implemented in Experiment 2 (i.e., dissociating FA rates through modality attributions) had the anticipated effect of decreasing FA rates in the mixed-aloud condition, whereas other aspects of participants' memory performance were largely unchanged. This result emerged in both the full and first-block analyses when mixed-list hits were scored leniently.

[19] When hits were scored stringently in Experiment 2, the ANOVA also revealed a significant main effect of Study List Type, $F(1, 175) = 33.02$, $MSE = 0.02$, $p < 0.001$, $\eta^2 = 0.16$, signifying a robust overall mixed-list cost of reading silently across both experiments. The main effect of experiment was nonsignificant, $F(1, 175) = 1.97$, $MSE = 0.02$, $p = 0.16$, $\eta^2 = 0.01$. More important, the Experiment × Study List Type interaction was significant, $F(1, 175) = 7.44$, $MSE = 0.02$, $p < 0.01$, $\eta^2 = 0.04$, reflecting the fact that the mixed-list cost of reading silently was larger in Experiment 2, $t(88) = 6.40$, $p < 0.001$, $d = 1.36$, than it was in Experiment 1, $t(87) = 2.01$, $p < 0.05$, $d = 0.43$. We contend that hit rates in the mixed-silent condition of Experiment 2 were particularly low due to the increased difficulty of making correct modality attributions.

[20] When stringently scored hit rates were used to derive the mixed-silent d' values in Experiment 2, the ANOVA yielded a significant main effect of study list type, $F(1, 175) = 12.77$, $MSE = 0.48$, $p < 0.001$, $\eta^2 = 0.07$, signifying that mixed-silent d' values tended to be lower, overall, than pure-silent d' values. The main effect of experiment was nonsignificant, $F(1, 175) = 1.04$, $MSE = 0.48$, $p = 0.31$, $\eta^2 = 0.01$. More important, there was a significant Experiment × Study List Type interaction, $F(1, 175) = 13.87$, $MSE = 0.48$, $p < 0.001$, $\eta^2 = 0.07$, reflecting the fact that there was a significant mixed-list d' cost in Experiment 2, $t(88) = 4.59$, $p < 0.001$, $d = 0.98$, but not in Experiment 1, $t(80.23) = 0.12$, *ns*.

(*Appendices continue*)

Appendix C

Comparing Experiment 3 with Bodner et al. (2014, Experiment 1)

By giving participants who studied a mixed list a 'pure' test (of either all aloud or all silent targets) we were able to dissociate mixed-list FA rates, which was not possible in Bodner and colleagues' (2014) Experiment 1: They used the overall mixed-list FA rate as an estimate for both mixed-aloud and mixed-silent FA rates. We predicted that dissociating FA rates in our experiment would lead to a significantly greater mixed-list d' benefit than that observed in Bodner and colleagues' Experiment 1. This was expected to arise due to mixed-aloud FA rates being lower in Experiment 3 than in Bodner and colleagues' Experiment 1. In contrast, we did not expect hit rates to differ significantly between these two experiments. To test these predictions, we conducted a series of two-way ANOVAs—on hits, FAs, and d'—in which study list type (mixed-aloud vs. pure-aloud) and experiment (our Experiment 3 vs. Bodner and colleagues' Experiment 1) were both between-subjects factors.

First, we conducted this ANOVA with hit rates as the dependent measure. This ANOVA revealed a main effect of study list type, $F(1, 204) = 11.06$, $MSE = 0.02$, $p = .001$, $\eta^2 = 0.05$, reflecting an overall benefit of reading aloud in a mixed versus pure list across both experiments. There was also a significant main effect of experiment, $F(1, 204) = 13.07$, $MSE = 0.02$, $p < .001$, $\eta^2 = 0.06$, consistent with there being higher overall hit rates in our Experiment 3 than in their Experiment 1. (This main effect was not surprising given that our test design change resulted in a recognition test that was half the length of that used in Experiment 1 of Bodner and colleagues, as outlined in our Experiment 3 method section). Of main interest, the Experiment × Study List Type interaction was not reliable, $F(1, 204) = 1.60$, $p = .21$, $MSE = 0.02$, $\eta^2 = .01$, suggesting that our "pure-test" manipulation did not influence the size of the mixed-list benefit in hits. There was a mixed-list benefit of reading aloud both in our Experiment 3, $t(110) = 1.78$, $p = .08$, $d = 0.34$, and a significant benefit in their Experiment 1, $t(94) = 2.73$, $p = .007$, $d = 0.56$.

In terms of FAs, the same two-way ANOVA yielded a nonsignificant main effect of study list type, $F(1, 204) = 2.70$, $MSE = 0.01$, $p = .10$, $\eta^2 = 0.01$. There was a reliable main effect of experiment, $F(1, 204) = 8.90$, $MSE = 0.01$, $p = .003$, $\eta^2 = 0.04$, signifying that the FA rates in these two conditions tended to be lower in our experiment than in theirs. More important, there was a significant Experiment × Study List Type interaction, $F(1, 204) = 20.00$, $MSE = 0.01$, $p < .001$, $\eta^2 = 0.09$. As anticipated, this interaction was driven by the fact that our data showed a mixed-list benefit of reading aloud in terms of lower FA rates, $t(100.82) = 2.22$, $p < .05$, $d = 0.42$. Bodner et al. (2014) found the opposite: lower FA rates in their pure-aloud condition than in their mixed-aloud condition, $t(80.07) = 3.91$, $p < .001$, $d = 0.80$.

Last, we conducted the same two-way ANOVA for d'. Again, the main effect of Study List Type was not reliable, $F(1, 204) = 2.19$, $MSE = 0.50$, $p = .14$, $\eta^2 = 0.01$, but there was a significant main effect of Experiment, $F(1, 204) = 21.15$, $MSE = 0.50$, $p < .001$, $\eta^2 = 0.09$, reflecting better memory performance in our experiment compared to theirs. More important, this ANOVA yielded a significant Experiment × Study List Type interaction, $F(1, 204) = 3.76$, $MSE = 0.50$, $p = .05$, $\eta^2 = 0.02$, reflecting the fact that there was a significant d' benefit of reading aloud in our Experiment 3, $t(110) = 2.54$, $p = .01$, $d = 0.48$, whereas there was no d' benefit in their Experiment 1, $t(94) = 0.31$, $p = .76$, $d = 0.06$.

Overall, this comparison between our data and the data of Experiment 1 in Bodner et al. (2014) suggests that our design modification (i.e., using pure tests after a mixed study list to yield separate mixed FA rates) enhanced the mixed-list benefit of reading aloud relative to an equivalent design in which a standard mixed recognition test was used following the mixed study list. Both experiments showed a mixed-list benefit in terms of increased hits, but only our Experiment 3 revealed a mixed-list benefit in terms of reduced FAs.

Conversely, combined analyses did not yield evidence that the mixed-list cost of reading silently was larger in our Experiment 3 than in Bodner and colleagues' (2014) Experiment 1. To compare the costs in these two experiments, we conducted a two-way ANOVA, in which study list type (mixed-silent vs. pure-silent) and experiment (our Experiment 3 vs. their Experiment 1) were both between-subjects factors.

In terms of hit rates, the ANOVA revealed a main effect of study list type, $F(1, 204) = 11.87$, $MSE = 0.02$, $p = .001$, $\eta^2 = 0.06$, reflecting an overall cost of reading silently in a mixed versus pure list across both experiments. There was also a significant main effect of experiment, $F(1, 204) = 14.89$, $MSE = 0.02$, $p < .001$, $\eta^2 = 0.07$, consistent with there being higher overall hit rates in our Experiment 3 than in their Experiment 1. There was not, however, a significant Experiment × Study List Type interaction ($F < 1$), suggesting that these costs did not differ in magnitude. Our Experiment 3 showed a mixed-list cost of reading silently, $t(99.68) = 2.22$, $p = .03$, $d = 0.42$, as did their Experiment 1, $t(94) = 2.62$, $p = .01$, $d = 0.54$.

In terms of FAs, the same two-way ANOVA yielded a nonsignificant main effect of study list type ($F < 1$). The main effect of experiment was also nonsignificant, as was the Experiment × Study List Type interaction (Fs < 1). Neither experiment showed a difference in FA rates between the mixed-silent and pure-silent conditions (both ts < 1). Thus, FA data in the silent conditions of these two experiments were very similar, with neither showing evidence of a mixed-list cost.

(*Appendices continue*)

Last, in terms of d', there was a significant main effect of Experiment, $F(1, 204) = 8.26$, $MSE = 0.52$, $p = .004$, $\eta^2 = 0.04$, consistent with better memory discrimination in our Experiment 3 than in their Experiment 1. There was also a significant main effect of study list type, $F(1, 204) = 5.80$, $MSE = 0.52$, $p = .02$, $\eta^2 = 0.03$, reflecting a modest overall mixed-list cost of reading silently. More importantly, the Experiment × Study List Type interaction was nonsignificant ($F < 1$), signifying that the size of the mixed-list cost did not differ reliably between the two experiments. There was a trend in the direction of a mixed-list cost of reading silently in our experiment, $t(104.53) = 1.47$, $p = .14$, $d = 0.28$, and a marginally significant cost in their experiment, $t(94) = 1.90$, $p = .06$, $d = 0.39$. Both effect sizes were modest.

In sum, a comparison of our Experiment 3 and Bodner and colleagues' (2014) Experiment 1 revealed that our "pure test" manipulation fostered a larger memorial benefit of reading aloud in a mixed study list. However, the cost of reading silently in a mixed list does not appear to have been influenced by our manipulation. We suspect that the benefit may have been accentuated in our Experiment 3 because participants in the mixed-aloud condition were aware that silent items would not appear on the recognition test. This may have increased participants' reliance on a speech distinctiveness heuristic due to its high diagnositicity: Participants could reject words on the basis of their lacking a record of speech—thereby lowering their FA rates—without the risk of incorrectly rejecting silent items.

Received June 4, 2014
Revision received September 7, 2015
Accepted September 29, 2015 ■