Reassessing the Basis of the Production Effect in Memory

Glen E. Bodner and Alexander Taikh
University of Calgary

The production effect refers to a memory advantage for items studied aloud over items studied silently. Ozubko and MacLeod (2010) used a list-discrimination task to support a distinctiveness account of the production effect over a strength account. We report new findings in this task—including negative production effects—that better fit with an attributional account of this task. According to the attributional account, list judgments are influenced by recognition memory, knowledge of the composition of the 2 lists, and a bias to attribute non-recognized items to the 1st list. Using a recognition task to eliminate these attributional influences revealed production effects consistent with either a distinctiveness or strength account. In our discussion, we consider whether the absence of production effects on implicit-memory tests and in between-group designs provides unequivocal support for a distinctiveness account over a strength account.

Keywords: production effect, list-discrimination task, recognition task, distinctiveness account, strength account

Becoming more active in the encoding of an item can enhance memory for it. For example, generating a word from a cue often sponsors better memory than simply reading it (e.g., Slamecka & Graf, 1978), and performing an action in response to an instruction results in better memory than watching an experimenter perform that action (e.g., Engelkamp, 1998). Recently, MacLeod, Gopie, Hourihan, Neary, and Ozubko (2010) confirmed Hopkins and Edwards’s (1972) neglected finding that simply saying items aloud at encoding can produce a memory benefit over reading them silently. Macleod et al. dubbed this benefit the production effect. They noted that the lack of research about this effect is surprising given that it is comparable in potency to the generation and enactment effects described above. Moreover, because it is simple and easy to implement, production might have the potential to be a useful study strategy. The production effect and explanations for it are the focus of the present article.

In their initial demonstration, Hopkins and Edwards (1972) tested a within-subjects group, who studied half of a list of words silently and half aloud, and two between-subject groups, who studied all the words either aloud or silently. On recognition and two-alternative forced choice (2AFC) tests, the within-subject group showed a recognition advantage for Aloud over Silent words, whereas no production effect occurred across the Aloud and Silent between-subject groups. Hopkins and Edwards concluded that for production to benefit memory “the study list must be mixed with respect to pronunciation” and that “the pronunciation effects are study-trial or encoding effects” (p. 536). Conway and Gathercole (1987) and Gathercole and Conway (1988) replicated the within-subjects production effect on recognition, and they also attributed the effect to an encoding locus. Specifically, they suggested that saying words aloud enhanced their relative distinctiveness in memory. MacDonald and MacLeod (1998) reported another two replications.

This set of four studies constituted the entire literature on production effects on explicit memory tasks until MacLeod and colleagues named and further delineated this effect (Hourihan & MacLeod, 2008; MacLeod, 2011; MacLeod et al., 2010; Ozubko, Gopie, & MacLeod, 2012; Ozubko & MacLeod, 2010).1 In a key article, MacLeod et al. (2010) provided several replications of the recognition advantage for Aloud over Silent words, and they identified some boundary conditions as well. For example, making the same overt response to all Aloud words (saying “yes” or pressing the space bar) did not lead to a production effect, but mouthing each word silently did so, showing that the response to each item had to be unique rather than having to be said aloud per se (see also MacLeod, 2011). In addition, MacLeod et al. consistently found production effects using a within-subjects design but not using a between-subject design (see also Hopkins & Edwards, 1972; cf. Gathercole & Conway, 1988, Experiment 5, however).

As reviewed in MacLeod et al. (2010), both decision-based and memory-based accounts of these production effects are tenable.

1 Dodson and Schacter (2001) are often cited as providing another demonstration of the production effect. However, they compared “Say” conditions to “Hear” (cf. Silent) conditions in which participants saw the items on the screen and heard them via headphones. Conway and Gathercole (1987, Experiment 4) and Gathercole and Conway (1988, Experiments 1, 2, and 5) consistently found better recognition in Heard conditions than Silent conditions (i.e., hearing words produces its own memory advantage). Because Dodson and Schacter’s design likely underestimated the benefits of production, we do not list it as a test of the production effect.
Decision-based accounts focus on participants’ metamemorial beliefs at test about what types of information they should expect to remember from study, whereas memory-based accounts focus on participants forming stronger memory traces for some items relative to others at the time of study (McCabe, Presmanes, Robertson, & Smith, 2004). The distinctiveness account is an example of a decision-based account (e.g., Dodson & Schacter, 2001). MacLeod et al. (2010) suggested that Aloud items are distinctive because they are “differentiated by being processed distinctively against the backdrop of silently read (unproduced) words” (p. 681). Recollecting having said an item aloud during study can be used heuristically at test to guide memory decisions (i.e., “if I remember saying it aloud then it must have been studied”). By this account, between-subject groups do not show a production effect because all words were studied in the same manner, thus Aloud words are not rendered distinct relative to another class of words (e.g., Hunt, 2006), and thus participants do not adopt a distinctiveness heuristic at test.

In contrast, the strength account provides a classic memory-based explanation for the production effect (Ozubko & MacLeod, 2010). Simply put, stronger memory traces are formed for Aloud words than Silent words, thus Aloud words should outperform Silent words on memory tasks. The strength account thus does not accord with the absence of a between-subject production effect, leading MacLeod et al. (2010) and Ozubko and MacLeod (2010) to suggest that the distinctiveness account of the production effect is preferable to the strength account.

In an attempt to provide a direct test of these two accounts, Ozubko and MacLeod (2010) designed a list-discrimination task, which also set off our initial experiments. In this task, participants study two lists of words and then complete a list-discrimination test in which they must assign each studied word to one of the two lists. In their Experiment 2, Aloud-Mixed and Silent-Mixed groups were tested. For both groups, half of the List 2 words were Aloud, and half were Silent; List 1 was exclusively Aloud or Silent items, respectively. Critically, a production effect occurred on List 2 in the Mixed-Silent group: The proportion of accurate list judgments was greater for Aloud than Silent items. In contrast, no difference between Silent and Aloud items occurred on List 2 in the Aloud-Mixed group. Their Experiment 1 replicated this pattern when List 1 was the Mixed list.

Ozubko and MacLeod (2010) took this pattern of results as support for the distinctiveness account over the strength account. In the Silent-Mixed group, a distinctiveness heuristic is useful at test because recollecting that an item was read aloud is diagnostic of list source, given that only List 2 contained Aloud items. However, in the Aloud-Mixed group, use of this form of a distinctiveness heuristic should be abandoned because recalling that an item was read aloud is not diagnostic of list source, given that both lists contained Aloud items. In contrast, the strength account predicts a production effect in both groups, because Aloud items should have been more strongly encoded than Silent items, regardless of the composition of the two lists. Ozubko and MacLeod concluded that the production effect results from the use of a distinctiveness heuristic: Saying items aloud at study makes them distinctive, and recollection of this information can be used strategically at test to improve memory accuracy.

We set out to replicate and then further test the distinctiveness and strength accounts of the production effect using the list-discrimination task. Experiment 1 was designed as a replication of Ozubko and MacLeod (2010, Experiment 2). To foreshadow, our results did not completely accord with theirs, and in fact, they challenged both of these accounts. In reflecting on our results, we realized that the list-discrimination task—because it requires participants to assign all studied words to either List 1 or List 2, regardless of whether they recollect which list they were on—is prone to attributional biases and inferences, as is true of source-monitoring tasks in general (e.g., Johnson, Hashtrudi, & Lindsay, 1993).

Within Experiment 1, we describe an attributional account of the production effect in the list-discrimination task. We then provide further support for this account of the list-discrimination task in Experiments 2 and 3. Finally, Experiment 4 shows that switching to a recognition task eliminates these attributional influences and reinstates a pattern of results that is consistent with either a strength or distinctiveness account. We conclude that the list-discrimination task is not useful for testing the strength and distinctiveness accounts of the production effect. Our General Discussion focuses on whether other evidence in the literature on the production effect unequivocally supports a distinctiveness account over a strength account.

**Experiment 1:**

**Silent-Mixed Versus Aloud-Mixed in List Task**

Experiment 1 tested Silent-Mixed and Aloud-Mixed groups, as in Ozubko and MacLeod (2010, Experiment 2). We expected to replicate their findings: a production effect on List 2 in the Silent-Mixed group, and a null difference on List 2 in the Aloud-Mixed group, consistent with the distinctiveness account. The strength account, in contrast, predicts production effects regardless of list composition. For replication purposes, a Mixed-Aloud group was subsequently tested, as in Ozubko and MacLeod (2010, Experiment 1).

**Method**

**Participants.** Sixteen University of Calgary undergraduates participated for course credit in each reported group in each experiment; no participant was tested twice. Random assignment to the Silent-Mixed or Aloud-Mixed group was used.

**Stimuli and apparatus.** The critical stimuli in each experiment were 64 nouns, 5–10 letters each, selected from the appendix of MacDonald and MacLeod (1998); the stimuli in Ozubko and MacLeod (2010) were also taken from this set. Stimuli and instructions were presented on a computer.

**Procedure.** Participants were tested individually. The 64 words were assigned to two lists of 32 words each; this assignment was counterbalanced across subjects. Participants were informed that they would study two lists of words for an unspecified memory test. An instruction screen labeled “List 1” or “List 2” preceded each list, informing subjects how they would study the words on that list. Depending on the task for a given list, subjects were told to read the blue words silently and/or to read the red words aloud. For Aloud lists, all 32 words were red. For Silent lists, all 32 words were blue. For Mixed lists, 16 words were blue (Silent), and 16 were red (Aloud); color assignment to items was counterbalanced across participants. Words were shown on the screen for 2 s, with a blank 0.5-s interstimulus interval, in 36-point Arial font.
After studying the two lists, participants received the list-discrimination test. They were shown the 64 studied words one at a time on the screen in a freshly randomized order. The test items were presented in black, again in 36-point Arial font. Participants responded using a button box. They pressed the left button (labeled “L1”) if they thought the word had appeared on List 1, or they pressed the right button (labeled “L2”) if they thought the word had appeared on List 2.

**Results and Discussion**

The mean proportions of correct list judgments for Silent and Aloud words on List 2 in the Silent-Mixed and Aloud-Mixed groups were compared using a 2 (group: Silent-Mixed, Aloud-Mixed) × 2 (item type: Aloud, Silent) mixed-factor analysis of variance (ANOVA; see Table 1). Results were significant at the .05 level unless otherwise noted. The group effect was not significant; accuracy was similar in the Silent-Mixed and Aloud-Mixed groups (.70 vs. .65), F(1, 30) = 1.05, MSE = .04, p = .31. The item-type effect was marginal, F(1, 30) = 3.17, MSE = .01, p = .09, but was qualified by a significant interaction, F(1, 30) = 40.32, MSE = .01.

The first follow-up test of this interaction revealed a production effect in the Silent-Mixed group (Silent = .58, Aloud = .83), F(1, 15) = 27.66, MSE = .02, replicating Ozubko and MacLeod (2010). This production effect fits with both the strength account (i.e., Aloud words had stronger memory traces) and the distinctiveness account (i.e., participants attempted to recollect whether a word was read aloud because this was diagnostic of list source given that only List 2 contained Aloud words).

The second follow-up test revealed an unexpected negative production effect in the Aloud-Mixed group (Silent = .72, Aloud = .58), F(1, 15) = 12.97, MSE = .01, in contrast to the null difference that Ozubko and MacLeod (2010) reported (Silent = .64, Aloud = .62). The strength account, of course, predicted a production effect. The distinctiveness account, in contrast, predicted no systematic difference between Aloud and Silent words. Use of a distinctiveness heuristic should have been turned off in this group because recollecting that a word was studied aloud is not indicative of list source. Neither account explains this negative effect.

Our failure to replicate Ozubko and MacLeod’s (2010, Experiment 2) null effect in the Aloud-Mixed group is surprising given that we used their materials and followed their procedure. Therefore, we tested a Mixed-Aloud group to follow-up this discrepancy (see Table 1). Ozubko and MacLeod (2010, Experiment 1) reported a nonsignificant negative production effect on List 1 (Silent = .62, Aloud = .56). In contrast, our Mixed-Aloud group produced a large negative production effect (Silent = .73, Aloud = .58), F(1, 15) = 7.68, MSE = .02, similar in magnitude to that obtained in our Aloud-Mixed group. In conjunction with an additional negative production effect in our Experiment 2, we suggest Ozubko and MacLeod’s null effects were due to subtle differences in their test instructions or procedure.

Whatever their locus, the occurrence of negative production effects strongly suggests that list judgments can be influenced by sources of information other than an item’s strength or recollecting having said it aloud. Although these negative production effects cannot be reconciled with a strength account, they might be reconciled within a distinctiveness account. Specifically, the Aloud-Mixed group may have taken their failure to recollect saying the Silent items aloud as evidence that they were from List 2 (or, for the Mixed-Aloud group, from List 1). By this account, the Aloud-Mixed and Mixed-Aloud groups applied the same distinctiveness heuristic used by the Silent-Aloud group, rather than “shutting it off,” because its failure would be at least suggestive of an item’s list source.

Ozubko and MacLeod (2010) did not endorse this version of the distinctiveness account because failing to recollect saying an item aloud could mean either that it was a Silent item or that it was an Aloud item that the participant failed to recollect having said aloud. However, Ozubko and MacLeod acknowledged the possibility that participants might adopt an “absence of recollection” decision rule, even if it were not perfectly diagnostic. Indeed, Dodson and Schacter (2001) provided evidence for use of such a rule: False recognition in the Deese–Roediger–McDermott (DRM) paradigm was reduced for critical lures from both Aloud lists and Silent lists in a within-subject design, suggesting participants took the absence of recollecting having said the critical lures from both types of list aloud as diagnostic evidence that they were not studied.

An alternative possibility is that our participants did not typically attempt to recollect saying or not saying the items aloud, nor even which list they were on. Indeed, we will suggest that their list judgments were based on a less effortful set of attributional biases. According to our attributional account of the list-discrimination task, participants infer the list source of an item based on their recognition experience for it, combined with other “handy” biases including their knowledge of the composition of each list (Bink, Marsh, & Hicks, 1999). The reliable occurrence of within-subjects production effects in the recognition task (e.g., MacLeod et al., 2010) means, de facto, that participants will be more likely to recognize Aloud words than Silent words. The attributional account assumes that there is a genuine recognition advantage for Aloud words, which could reflect either greater strength or distinctiveness for Aloud words. Either way, the Silent-Mixed group, as long as they realize that only List 2 contained Aloud words, should be biased to attribute recognized items to List 2, consistent with the observed production effect. In contrast, the Aloud-Mixed group, as long as they realize that only List 2 contained Silent words, should be biased to attribute non-recognized items to List 2, consistent with the observed negative production effect. Use of a list-composition inference or bias thus accommodates the Ex-

### Table 1

**Experiments 1–2: Mean Proportion of Correct List Judgments for Words From Each Study List**

<table>
<thead>
<tr>
<th>Experiment/Group</th>
<th>List 1</th>
<th>List 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Experiment 1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Silent-Mixed</td>
<td>.68 (.04)</td>
<td>.83 (.02)</td>
</tr>
<tr>
<td>Aloud-Mixed</td>
<td>.63 (.03)</td>
<td>.58 (.05)</td>
</tr>
<tr>
<td>Mixed-Aloud</td>
<td>.58 (.04)</td>
<td>.73 (.03)</td>
</tr>
<tr>
<td>Experiment 2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mixed-Mixed</td>
<td>.56 (.04)</td>
<td>.68 (.03)</td>
</tr>
</tbody>
</table>

*Note.* Standard errors are in parentheses.
Experiment 1 findings without positing that participants use a distinctiveness heuristic.

Experiment 2: Mixed-Mixed in List Task

Experiment 2 was designed to wedge apart the predictions of the strength, distinctiveness, and attributional accounts of the Experiment 1 findings. To this end a Mixed-Mixed group was tested: Half the items on each list were Silent, and half were Aloud. The strength account predicts production effects on both lists. On either version of the distinctiveness account described above, because both lists contain both types of items, neither recollecting nor failing to recollect saying items aloud will provide useful list-source information. Therefore, use of either type of distinctiveness heuristic should be turned off, resulting in equivalent accuracy on Silent and Aloud items within each list.

According to the attributional account, the Mixed-Mixed group would need to rely on a basis other than list composition for making list attributions. We suggest that in the list-discrimination task, participants are biased to attribute non-recognized words to the earlier list. Given that Aloud words are more likely to be recognized than Silent words (i.e., given that the production effect exists), participants should be biased to attribute non-recognized items to List 1 (and/or to attribute recognized items to List 2). This list-order bias should lead to a production effect on List 2 (by boosting accuracy for Silent items relative to Aloud items) and a negative production effect on List 1 (by reducing accuracy for Silent items relative to Aloud items).

Method

The method was the same as in Experiment 1 except here both lists were Mixed lists.

Results and Discussion

A 2 (list: List 1, List 2) × 2 (item type: Silent, Aloud) repeated-measures ANOVA was conducted on correct list judgments (see Table 1). The main effect of list was not significant; accuracy was similar for List 1 and List 2 items (.62 vs .59), $F < 1$. The main effect of item type was also not significant; accuracy was similar for Silent and Aloud items averaged across list (.58 vs .62), $F(1, 15) = 1.87, MSE = .02, p = .19$. However, the anticipated interaction was highly significant, $F(1, 15) = 26.97, MSE = .02$. Follow-up tests revealed a significant negative production effect on List 1 (Silent = .68, Aloud = .56), $F(1, 15) = 6.21, MSE = .02$, and a significant production effect on List 2 (Silent = 0.48, Aloud = 0.69), $F(1, 15) = 21.38, MSE = .02$.

The presence of both a production effect and a negative production effect in the Mixed-Mixed group contradicts both the strength account (which predicted production effects on both lists) and the distinctiveness account (which predicted null effects on both lists). This pattern fits well with the attributional account of the list-discrimination task, under the assumption that participants were biased to attribute non-recognized items to the earlier list (i.e., a list-order bias). In the General Discussion, we link these effects to bias effects in the source judgment literature (e.g., Bink et al., 1999).

Experiment 3: Aloud-Mixed and Mixed-Mixed in List Task With Lures

The goal of Experiment 3 was to provide an additional test of the attributional account of the list-discrimination task. We suggest that the negative production effects in Experiments 1 and 2 were due to the forced-response nature of the list-discrimination task. When participants must classify each test item as having been on List 1 or List 2, then when they fail to recognize an item, they will be forced to rely on available heuristic biases such as list composition (Experiment 1) or list order (Experiment 2) that can result in negative production effects. In Experiment 3, we included new items on the test, and participants were given the option of classifying non-recognized items as “new.” We predicted that participants would now be biased to attribute Silent items to the “new” category, which would work to reduce accuracy for Silent items, thus undoing the negative production effects observed in Experiments 1 and 2. Experiment 3A tested an Aloud-Mixed group, and Experiment 3B tested a Mixed-Mixed group.

Method

The method was the same as in earlier experiments except as noted here. Additional words from MacDonald and MacLeod’s (1998) appendix served as the new items. At test, participants received the 64 studied words and 32 new items in a random order. They were instructed that the test consisted of words they had studied and words they had not studied. They pressed the left button (labeled “L1”) if they thought the word had appeared on List 1, the middle button (labeled “L2”) if they thought the word had appeared on List 2, and the right button (labeled “new”) if they thought the word had not appeared on either list.

Results and Discussion

Table 2 provides the mean proportions of correct list judgments, incorrect list judgments, and new judgments for studied words, and the mean proportion new words incorrectly assigned to List 1 or List 2 in Experiment 3.

In Experiment 3A, the Aloud-Mixed group showed a production effect on List 2 for correct list judgments (Silent = .47, Aloud = .57), $F(1, 15) = 5.36, MSE = .02$. On the attributional account, the negative production effect in the Aloud-Mixed group in Experiment 1 did not occur here because participants were now biased to attribute non-recognized items to the “new” category rather than to List 2. Indeed, “new” judgments were far more likely for Silent than Aloud items on List 2 (Silent = .36, Aloud = .16), $F(1, 15) = 12.70, MSE = .02$ (for a similar finding using Silent/Aloud judgments instead of List 1/List 2 judgments, see Ozubko et al., 2012, Experiment 3). There was no difference in the rate at which new items were incorrectly attributed to List 1 versus List 2 (.08 vs .10), $F < 1$. At the risk of over interpreting this null effect, we note that residual use of a list-composition bias would work to increase List 2 judgments for new items (because only List 2 contained Silent items), and residual use of a list-order bias would work to increase List 1 judgments for new items (because List 1 was presented earlier). Thus, it is possible that these two biases canceled each other out.
In Experiment 3B, correct list judgments in the Mixed-Mixed group (see Table 2) were analyzed as in Experiment 2. Accuracy was marginally lower on List 1 than List 2 (.45 vs. .53), $F(1, 15) = 4.44$, $MSE = .02$, $p = .052$. Most importantly, there was a robust production effect (Silent = .40, Aloud = .58), $F(1, 15) = 27.50$, $MSE = .02$, that did not interact with list, $F < 1$. On the attributional account, the negative production effect on List 1 in Experiment 2 did not occur here because participants were now biased to attribute non-recognized items to the new category rather than to List 1. Indeed, a 2 (list: List 1, List 2) × 2 (item type: Silent, Aloud) repeated-measures ANOVA on “new” list judgments revealed that far more Silent items than Aloud items were classified as new (.34 vs. .17), $F(1, 15) = .28$, $MSE = .02$, that did not interact with list, $F < 1$. The trend for List 1 items to receive more new judgments than List 2 items was not significant (.28 vs. .23), $F(1, 15) = 2.05$, $MSE = .02$, $p = .17$.

Finally, new items were more likely to be incorrectly assigned to List 1 than to List 2 (.12 vs. .07), $F(1, 15) = 6.35$, $MSE = .003$, consistent with a residual list-order bias in the Mixed-Mixed group, which would not be counteracted by a residual list-composition bias, given that the two lists were both composed of Aloud and Silent items. This finding bolsters our contention that the Mixed-Mixed group in Experiment 2 relied on a list-order bias. In the General Discussion, we suggest this finding is also consistent with an account of bias effects reported in the source-monitoring literature (e.g., Bink et al., 1999; Hoffman, 1997).

**Experiment 4: Aloud-Mixed and Mixed-Mixed in a Recognition Task**

Experiment 3 showed that the same conditions that led to negative production effects in Experiments 1 and 2 led to production effects when participants were presented with new items and were given a “new” response option. However, it is not clear to what extent the production effects in Experiment 3 reflected a memory-based advantage for Aloud items versus a bias to assign weakly recognized words to the “new” category. That is, participants may have erred on the “safe side” (or “lazy side”) by choosing to classify weakly recognized items as “new” rather than considering them to be candidate “Silent” items, because the latter would require them to have to determine on which of the two lists each Silent item had been presented. A bias to classify weakly recognized items as “new” would therefore work to underestimate accuracy for Silent items, thus exaggerating the magnitude of the production effect.

Given that even the modified list-discrimination task used in Experiment 3 remains susceptible to attributional biases, the goal of Experiment 4 was to establish that production effects do occur reliably in the two-study-list paradigm when these biases are not in play. To this end, Experiment 4 was a replication of Experiment 3 using a standard recognition task. We assume that recognition judgments would be based on participants’ recognition experiences alone, and that they would not spontaneously attempt to determine the list origin of items prior to making their responses. Thus, participants’ knowledge of the composition and order of the two lists were not expected to influence recognition judgments. As a result, we expected to obtain production effects in the Aloud-Mixed and Mixed-Mixed groups.

**Method**

The method was as in Experiment 3 except as noted here. Additional words from MacDonald and MacLeod’s (1998) appendix served as new items ($n = 46$; 29 in common with Experiment 3). The test presented the 64 studied words and the 46 new items in a random order. Participants were instructed that the test consisted of words they studied and words they had not studied. They pressed the left button (“old”) if they thought the word had appeared on one of the two lists at study, and they pressed the right button (“new”) if they thought the word had not been studied.

**Results and Discussion**

As expected, production effects in the two-study-list paradigm were obtained on a standard recognition task. In Experiment 4A, the Aloud-Mixed group showed a production effect on List 2 (Silent = .70, Aloud = .81), $F(1, 15) = 6.15$, $MSE = .02$. In

<table>
<thead>
<tr>
<th>Experiment/Item type</th>
<th>List 1</th>
<th>List 2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Aloud</td>
<td>Silent</td>
</tr>
<tr>
<td>Experiment 3A: Aloud-Mixed group</td>
<td></td>
<td></td>
</tr>
<tr>
<td>List items</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Correct list judgments</td>
<td>.49 (.04)</td>
<td>.57 (.04)</td>
</tr>
<tr>
<td>Incorrect list judgments</td>
<td>.28 (.02)</td>
<td>.27 (.04)</td>
</tr>
<tr>
<td>“New” judgments</td>
<td>.23 (.03)</td>
<td>.16 (.04)</td>
</tr>
<tr>
<td>New items</td>
<td>.08 (.02)</td>
<td>.10 (.02)</td>
</tr>
<tr>
<td>Experiment 3B: Mixed-Mixed group</td>
<td></td>
<td></td>
</tr>
<tr>
<td>List items</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Correct list judgments</td>
<td>.53 (.05)</td>
<td>.37 (.04)</td>
</tr>
<tr>
<td>Incorrect list judgments</td>
<td>.26 (.04)</td>
<td>.28 (.03)</td>
</tr>
<tr>
<td>“New” judgments</td>
<td>.21 (.06)</td>
<td>.35 (.05)</td>
</tr>
<tr>
<td>New items</td>
<td>.12 (.03)</td>
<td>.07 (.02)</td>
</tr>
</tbody>
</table>

Note. Standard errors are in parentheses.
Experiment 4B, the Mixed-Mixed group showed a production effect (Silent = .59, Aloud = .82), $F(1, 15) = 43.12, \text{MSE} = .01$, that did not interact with list, $F < 1$. As in Experiment 3B, performance in Experiment 4B was nonsignificantly worse on List 1 than List 2 (.67 vs. .73), $F(1, 15) = 2.67, \text{MSE} = .02, p = .12$. As discussed below, these effects are consistent with the elimination of attributional influences in standard recognition, but they are also consistent with both strength and distinctiveness accounts of the production effect.

**General Discussion**

MacLeod et al. (2010) established a number of important boundary conditions for obtaining the production effect, leading them to favor a distinctiveness account over a simple strength account. Ozubko and MacLeod (2010) subsequently used a list-discrimination task to provide a direct test of the distinctiveness account. Our study revealed that the list-discrimination task is not useful for evaluating the distinctiveness and strength accounts, due to its susceptibility to attributional biases. In Experiment 1, Aloud-Mixed and Mixed-Aloud groups showed a negative production effect. According to MacLeod et al.’s version of a distinctiveness account, these groups should have abandoned use of a distinctiveness heuristic because recalling having said an item aloud would not reveal which list it had been studied on. This effect can be reconciled with a distinctiveness account, however, by assuming that both groups used a heuristic in which failing to recollect saying an item aloud guided responses (see Dodson & Schacter, 2001). Using a Mixed-Mixed group, Experiment 2 ruled out this version of the distinctiveness account. In a Mixed-Mixed group, attempting to recollect if an item was said aloud, or attempting to recollect if an item was read silently, are both uninformative regarding list source, so use of either heuristic should have been turned off. Yet, the Mixed-Mixed group showed a negative production effect on List 1 coupled with a production effect on List 2.

We offered a simple attributional account of how participants perform the list-discrimination task. According to Jacoby, Kelley, and Dywan (1989), the familiarity experienced for a stimulus does not specify its source but “is attributed to a particular source depending on the details of the experimental situation” (p. 404). We suggest that participants in Experiment 1 used their knowledge of the composition of the two lists to infer list source as a function of whether they recognized an item. Thus, the Silent-Mixed group was biased to attribute recognized items to List 2 (the only list containing Aloud items), whereas the Aloud-Mixed group was biased to attribute non-recognized items to List 2 (the only list containing Silent items). When list-composition could not be used to infer list source, as with the Mixed-Mixed group in Experiment 2, participants were biased to attribute non-recognized items to the earlier list. This list-order bias led Silent items to be attributed to List 1, thus contributing to a production effect on List 2 and a negative production effect on List 1. In sum, the attributional account proposes that list judgments are influenced by recognition memory, knowledge of the composition of the lists, and a bias to attribute non-recognized items to the earlier list.

We are not suggesting participants never recollect having said an item aloud (or recollect not having said an item aloud), or that participants never adopt a distinctiveness heuristic—even in the list-discrimination task. Rather, our point is that the pattern of effects in Experiments 1 and 2 show that attributional biases in the list-discrimination task were potent enough to produce reversals in the observed effects sufficient to obscure the underlying memorial benefit of production in this task. In Experiment 4, switching to a standard recognition task revealed production effects in the same groups where negative production effects were seen in Experiments 1 and 2, demonstrating the robustness of attributional biases in the list task. In Experiment 3, adding new items at test and providing a “new” response option also resulted in production effects rather than negative production effects. Participants in Experiment 3 thus tended to attribute non-recognized items to the “new” category. Given that Silent items are less likely to be recognized that Aloud items, this tendency worked to decrease accuracy for Silent items relative to Aloud items.

We cannot rule out the possibility that the Aloud-Mixed and Mixed-Mixed groups in Experiment 3 used a distinctiveness heuristic in which they tried to recall having said an item aloud. Although such recollections would not diagnose list source (given that both lists contained Aloud items), they would at least allow participants to rule out a “new” response. Estimated recognition in Experiment 3 (i.e., the sum of incorrect and correct list judgments rows in Table 3) was consistently higher for Aloud than Silent conditions, consistent with the distinctiveness (or strength) account. However, both groups would still have had to use a different (in our view, attributional) basis for making either a “List 1” or “List 2” judgment.

Experiment 3 ruled out use of a “negative” distinctiveness heuristic (Dodson & Schacter, 2001), on the other hand. If the Aloud-Mixed group had based their decisions on the failure to recall having said items aloud, the absence of such recollections would have boosted accuracy for Silent items on List 2 (given that only List 2 contained Silent items), which would work against the production effect we obtained on List 2. Moreover, the absence of such recollections would not have been useful for making List 1 versus List 2 judgments in the Mixed-Mixed group. As MacLeod and colleagues (MacLeod et al., 2010; Ozubko & MacLeod, 2010) have noted, the failure to recollect saying an item aloud is generally not diagnostic; in Experiment 3, it could indicate either that it was a Silent item, a new item, or an Aloud item the participant failed to recollect having said aloud.

A reviewer suggested a slight variant of the distinctiveness account might explain the results of Experiment 1. On this variant, participants in the Aloud-Mixed group would not abandon use of a distinctiveness heuristic because recollecting that an item was said aloud would not be diagnostic of whether they recognized an item. This effect can be reconciled with a distinctiveness account, however, by assuming that both groups used a heuristic in which failing to recollect saying an item aloud guided responses (see Dodson & Schacter, 2001). Using a Mixed-Mixed group, Experiment 2 ruled out this version of the distinctiveness account. In a Mixed-Mixed group, attempting to recollect if an item was said aloud, or attempting to recollect if an item was read silently, are both uninformative regarding list source, so use of either heuristic should have been turned off. Yet, the Mixed-Mixed group showed a negative production effect on List 1 coupled with a production effect on List 2.

We offered a simple attributional account of how participants perform the list-discrimination task. According to Jacoby, Kelley, and Dywan (1989), the familiarity experienced for a stimulus does not specify its source but “is attributed to a particular source depending on the details of the experimental situation” (p. 404). We suggest that participants in Experiment 1 used their knowledge of the composition of the two lists to infer list source as a function of whether they recognized an item. Thus, the Silent-Mixed group was biased to attribute recognized items to List 2 (the only list containing Aloud items), whereas the Aloud-Mixed group was biased to attribute non-recognized items to List 2 (the only list containing Silent items). When list-composition could not be used to infer list source, as with the Mixed-Mixed group in Experiment 2, participants were biased to attribute non-recognized items to the earlier list. This list-order bias led Silent items to be attributed to List 1, thus contributing to a production effect on List 2 and a negative production effect on List 1. In sum, the attributional account proposes that list judgments are influenced by recognition memory, knowledge of the composition of the lists, and a bias to attribute non-recognized items to the earlier list.

We are not suggesting participants never recollect having said an item aloud (or recollect not having said an item aloud), or that participants never adopt a distinctiveness heuristic—even in the list-discrimination task. Rather, our point is that the pattern of effects in Experiments 1 and 2 show that attributional biases in the list-discrimination task were potent enough to produce reversals in the observed effects sufficient to obscure the underlying memorial benefit of production in this task. In Experiment 4, switching to a standard recognition task revealed production effects in the same groups where negative production effects were seen in Experiments 1 and 2, demonstrating the robustness of attributional biases in the list task. In Experiment 3, adding new items at test and providing a “new” response option also resulted in production effects rather than negative production effects. Participants in Experiment 3 thus tended to attribute non-recognized items to the “new” category. Given that Silent items are less likely to be recognized that Aloud items, this tendency worked to decrease accuracy for Silent items relative to Aloud items.

We cannot rule out the possibility that the Aloud-Mixed and Mixed-Mixed groups in Experiment 3 used a distinctiveness heuristic in which they tried to recall having said an item aloud. Although such recollections would not diagnose list source (given that both lists contained Aloud items), they would at least allow participants to rule out a “new” response. Estimated recognition in Experiment 3 (i.e., the sum of incorrect and correct list judgments rows in Table 3) was consistently higher for Aloud than Silent conditions, consistent with the distinctiveness (or strength) account. However, both groups would still have had to use a different (in our view, attributional) basis for making either a “List 1” or “List 2” judgment.

Experiment 3 ruled out use of a “negative” distinctiveness heuristic (Dodson & Schacter, 2001), on the other hand. If the Aloud-Mixed group had based their decisions on the failure to recall having said items aloud, the absence of such recollections would have boosted accuracy for Silent items on List 2 (given that only List 2 contained Silent items), which would work against the production effect we obtained on List 2. Moreover, the absence of such recollections would not have been useful for making List 1 versus List 2 judgments in the Mixed-Mixed group. As MacLeod and colleagues (MacLeod et al., 2010; Ozubko & MacLeod, 2010) have noted, the failure to recollect saying an item aloud is generally not diagnostic; in Experiment 3, it could indicate either that it was a Silent item, a new item, or an Aloud item the participant failed to recollect having said aloud.

A reviewer suggested a slight variant of the distinctiveness account might explain the results of Experiment 1. On this variant, participants in the Aloud-Mixed group would not abandon use of a distinctiveness heuristic because recollecting that an item was said aloud is a more diagnostic characteristic of List 1 items (all of

<table>
<thead>
<tr>
<th>Experiment/Group</th>
<th>List 1</th>
<th>List 2</th>
<th>New items</th>
</tr>
</thead>
<tbody>
<tr>
<td>Experiment 4A</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Aloud-Mixed group</td>
<td>.78 (.04)</td>
<td>.81 (.03)</td>
<td>.70 (.04)</td>
</tr>
<tr>
<td>Experiment 4B</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mixed-Mixed group</td>
<td>.77 (.03)</td>
<td>.57 (.05)</td>
<td>.86 (.04)</td>
</tr>
</tbody>
</table>

Note. Standard errors are in parentheses.
which were Aloud items) than List 2 items (only half of which were Aloud items). Therefore, the Aloud-Mixed group would tend to attribute Aloud words to List 1 rather than List 2, in line with the negative production effect on List 2; the same type of argument also fits the Mixed-Aloud group results. Similarly, the Silent-Mixed group could take the failure to recollect that an item was said aloud as support for a List 1 judgment (given all List 1 items were Silent) rather than a List 2 judgment (given only half the List 2 items were Silent). This use of distinctiveness would lead the Silent-Mixed group to attribute Silent words to List 1 rather than List 2, in line with the production effect on List 2.

However, this version of the distinctiveness account cannot explain the results of Experiments 2 and 3. The Mixed-Mixed groups in Experiments 2 and 3 received two mixed lists, and the percentage of Silent and Aloud words on each list was identical, hence there was no diagnostic value to either recollecting saying a word aloud or failing to recollect saying a word aloud. Null effects should have resulted on each list, contrary to what was observed. Finally, if the Aloud-Mixed group based their list judgments on the relative purity of the two lists, then the negative production effect in the Aloud-Mixed group in Experiment 1 should also have occurred in the Aloud-Mixed group in Experiment 3. Instead, we showed that adding a “new” judgment option (and lure items) in Experiment 3 now brought about a production effect.

The strength account fared no better at accommodating our findings. For example, it consistently predicted production effects in Experiments 1 and 2. Furthermore, neither the strength nor distinctiveness accounts explain why Silent words were more likely than Aloud words to garner “new” responses in Experiment 3, nor for the observed patterns of List 1 and List 2 judgments to new items.

In contrast to these difficulties, both the strength and distinctiveness accounts accord well with the production effects in the standard recognition task in Experiment 4, thus bringing us full circle to the original problem of teasing apart these two accounts. Given that the list-discrimination task does not appear to be useful for this purpose, and given that both accounts are consistent with the occurrence of within-subject production effects in explicit-memory tasks, are there any other findings to date that might be useful for distinguishing the distinctiveness and strength accounts of the production effect? MacLeod et al. (2010) reported two other potentially relevant findings, which we next consider in turn.

The Absence of a Production Effect in an Implicit Speed-Reading Task

In MacLeod et al.’s (2010) experiments, the recognition/2AFC task was preceded by an implicit speed-reading task, consisting of Aloud (or other overt response condition) items and Silent items from the study phase, plus new items. Participants read these items aloud as quickly as possible. Although prior study reduced naming times in six of eight experiments, there was no production (or other production-type-task) effect (see also MacDonald & MacLeod, 1998). MacLeod et al. (see also Ozubko et al., 2012) interpreted these data as support for the distinctiveness account in that only on explicit-memory tests should participants intentionally attempt to recollect whether they had said items aloud at study, and hence only on explicit tests should a production effect occur. The absence of a production effect on an implicit task thus fits well with the distinctiveness account. Furthermore, it also rules out a version of a strength account we term an automatic-strength account in which stronger memory traces formed for Aloud items automatically spawn better memory task performance. But, on a more sophisticated version of the strength account that we term an evaluated-strength account, whether a production effect occurs on a task will depend on whether participants engage in intentional evaluation of the memory strength of items. On an implicit test such as speeded naming, participants are very unlikely to evaluate the strength of their memory for each item prior to making a response to it. Participants are not likely evaluating either qualitative aspects or the strength of their memories for items on implicit tests. If so, then although MacLeod et al.’s (2010) implicit task results rule out an automatic-strength account, they are consistent with either an evaluated-strength account or a distinctiveness account.

The Absence of a Production Effect in a Between-Subjects Design

To date, one recognition experiment has shown a between-subjects production effect (Gathercole & Conway, 1988, Experiment 5), and four have not (Hopkins & Edwards, 1972, Experiments 1 and 2; MacLeod et al., 2010, Experiments 2 and 3). Assuming the production effect is at least reduced (if not eliminated) in between-subject designs, this finding rules out an automatic-strength account described above. Critically, it would also seem to rule out an evaluated-strength account because participants are presumably evaluating memory strength information when making recognition decisions (cf. speeded naming responses), and memory strength should be greater for Aloud items than Silent items regardless of list composition. Thus, the absence (or reduction) of the production effect in a between (vs. within) design provides key support for a distinctiveness account over a strength account.

On the other hand, it is worth noting that the existence of between-subject production effects would not rule out a distinctiveness account. Instead, it would merely suggest that participants in pure Aloud groups adopt a distinctiveness heuristic strategy at test under some conditions. Speculatively, we suggest that participants may be unlikely to adopt a distinctiveness heuristic at test when presented with a production task on every trial at study (as opposed to a more obvious encoding task such as generation; e.g., Gunter, Bodner, & Azad, 2007). The production task may be interpreted solely as a manipulation check that they have encoded the items, rather than as an encoding task per se. In contrast, when some of the studied items are read silently, within-subject participants may be more likely to deem the act of production to be part of the experiment, and this awareness may lead them to adopt a distinctiveness rule at test. If so, making the production task more salient, or coaching participants to adopt a distinctiveness rule at test (as suggested by MacLeod et al., 2010), may yield a between-subject production effect. We are currently exploring such possibilities.

Recently, Ozubko et al. (2012) reported that production increases both recollection and familiarity as assessed using either remember/know judgments (Experiment 1) or the receiver-operating characteristics procedure (Experiment 2). Although
these results were not taken as support a distinctiveness account over a strength account, a strength-based signal detection model could only fit the data from both experiments if an odd assumption was made. Specifically, it had to be assumed that the Aloud and Silent item distributions were equally variable in Experiment 1, but that the Aloud item distribution was more variable than the Silent item distribution in Experiment 2. This possibility was deemed unlikely, given that both experiments used the same study phase and differed only in the response options offered at test. Ozubko et al. took these findings as support for a dual-process account of the production effect (i.e., separate recollection and familiarity processes), rather than support for a distinctiveness account, but they also appear to provide another challenge to the strength account. On balance, then, the extant literature better supports a distinctiveness account than a strength account of the production effect, even with the exclusion of the list-discrimination task findings.

Implications for the Origin of “It-Had-To-Be-You/It-Had-To-Be-Me” Source-Monitoring Effects

Although it was not the focus of our study, our findings help specify the basis of “it-had-to-be-you” and “it-had-to-be-me” biases in source-monitoring studies. Johnson, Raye, Foley, and Foley (1981) first reported the former type of bias: New items were more likely to be attributed to an external/perceived “read” source than to an internal/imagined “generate” source. Later, Hoffman (1997) first showed the reverse bias using pictures: New pictures were more likely to be attributed to an internal/imagined source than to an external/perceived picture source. In an interesting parallel to the present study, Hoffman designed further experiments to test between a strength account and a qualitative-characteristics (source-monitoring) account. Picture-imagined (PI) and imagined-picture (IP) groups were tested, referring to the respective sources on Day 1 and Day 3. The source-monitoring test occurred on Day 3. Consistent with the strength account, participants tended to attribute new items to the weaker source. Thus, the PI group generated an it-had-to-be-you effect, and the IP group generated an it-had-to-be-me effect. Hoffman argued that the qualitative-characteristics account, in contrast, predicted an it-had-to-be-me effect for both groups. Given that perceptual details should largely be absent for new items, they should have been attributed systematically to the imagined source in each group. Bink et al. (1999) challenged Hoffman’s (1997) interpretation by showing the same pattern of attributions for new items obtained by Hoffman occurred even when the items shown on Day 1 were presented three times and no longer constituted a weaker source (indeed, it was now the stronger source). Bink et al. argued in favor of the qualitative-characteristics account. Specifically, they claimed that the IP group’s attributions were driven by attempts to recollect perceptual details because they had just seen pictures prior to the source test, whereas the PI group’s attributions were driven by attempts to recollect cognitive operations details because they had just imagined items prior to the source test. They concluded that participants ascribe new items to the source associated with the least diagnostic memory characteristics, akin to the list-order bias we suggest operates in the list-discrimination task.

Although Bink et al.’s (1999) findings clearly challenge Hoffman’s (1997) strength account, they were unable to adjudicate between two qualitative-characteristics-based explanations. Were their participants biased to attribute new items to the Day 1 (vs. Day 3) source “because of an implicit belief that those items would be worse remembered or because there is more variability in the quality of what is retrieved” (Bink et al., 1999, p. 808)? Our results provide clear support for the former possibility. In Experiment 2, the Mixed-Mixed group showed a production effect on List 2 and a negative production effect on List 1. This pattern fits with Bink et al.’s claim that participants are biased to attribute non-recognized items to the earlier source (i.e., list) because of an implicit belief that that source would be remembered less well (i.e., a list-order bias). More dramatically, in Experiment 3, our Mixed-Mixed group was more likely to attribute new items to List 1 than to List 2, even though both lists contained both Silent and Aloud words. The Mixed-Mixed group clearly relied on a list-order bias. Critically, given that List 1 and List 2 were presented back-to-back in our experiments, in contrast to the 2-day delay in Bink et al.’s and Hoffman’s studies, it is highly unlikely that there would be more variability in the qualitative details recollected for List 1 items than for List 2 items. Given the absence of a significant difference in recognition accuracy for List 1 and List 2 items in the Mixed-Mixed group in Experiment 4, our results therefore suggest the operation of a mistaken metacognitive belief about the memorability of two lists presented in short succession. As such, the list-discrimination task, though it was not useful for separating strength and distinctiveness accounts of the production effect, contributes to our understanding of the basis of source-monitoring biases.

References

Call for Nominations

The Publications and Communications (P&C) Board of the American Psychological Association has opened nominations for the editorships of Behavioral Neuroscience, Journal of Applied Psychology, Journal of Educational Psychology, Journal of Personality and Social Psychology: Interpersonal Relations and Group Processes, Psychological Bulletin, and Psychology of Addictive Behaviors for the years 2015–2020. Mark S. Blumberg, PhD, Steve W. J. Kozlowski, PhD, Arthur Graesser, PhD, Jeffry A. Simpson, PhD, Stephen P. Hinshaw, PhD, and Stephen A. Maisto, PhD, respectively, are the incumbent editors.

Candidates should be members of APA and should be available to start receiving manuscripts in early 2014 to prepare for issues published in 2015. Please note that the P&C Board encourages participation by members of underrepresented groups in the publication process and would particularly welcome such nominees. Self-nominations are also encouraged.

Search chairs have been appointed as follows:

- Behavioral Neuroscience, John Disterhoft, PhD
- Journal of Applied Psychology, Neal Schmitt, PhD
- Journal of Educational Psychology, Neal Schmitt, PhD, and Jennifer Crocker, PhD
- Journal of Personality and Social Psychology: Interpersonal Relations and Group Processes, David Dunning, PhD
- Psychological Bulletin, Norman Abeles, PhD
- Psychology of Addictive Behaviors, Jennifer Crocker, PhD, and Lillian Comas-Diaz, PhD

Candidates should be nominated by accessing APA’s EditorQuest site on the Web. Using your Web browser, go to http://editorquest.apa.org. On the Home menu on the left, find “Guests.” Next, click on the link “Submit a Nomination,” enter your nominee’s information, and click “Submit.”

Prepared statements of one page or less in support of a nominee can also be submitted by e-mail to Sarah Wiederkehr, P&C Board Search Liaison, at swiederkehr@apa.org.

Deadline for accepting nominations is January 11, 2013, when reviews will begin.