Memory, 2014
Vol. 22, No. 5, 509–524, http://dx.doi.org/10.1080/09658211.2013.800554

Remembered study mode: Support for the distinctiveness account of the production effect

Jason D. Ozubko[1], Jennifer Major[2], and Colin M. MacLeod[3]

[1]Rotman Research Institute, Baycrest Centre, Toronto, Ontario, Canada
[2]Department of Psychology, Wilfrid Laurier University, Waterloo, Ontario, Canada
[3]Department of Psychology, University of Waterloo, Ontario, Canada

The production effect is the finding that words spoken aloud at study are subsequently remembered better than are words read silently at study. According to the distinctiveness account, aloud words are remembered better because the act of speaking those words aloud is encoded and later recovery of this information can be used to infer that those words were studied. An alternative account (the strength-based account) is that memory strength is simply greater for words read aloud. To discriminate these two accounts, we investigated study mode judgements (i.e., ‘‘aloud’’/’’silent’’/’’new’’ ratings): The strength-based account predicts that ‘‘aloud’’ responses should positively correlate with memory strength, whereas the distinctiveness account predicts that accuracy of study mode judgements will be independent of memory strength. Across three experiments, where the strength of some silent words was increased by repetition, study mode was discriminable regardless of strength—even when the strength of aloud and repeated silent items was equivalent. Consistent with the distinctiveness account, we conclude that memory for ‘‘aloudness’’ is independent of memory strength and a likely candidate to explain the production effect.

Keywords: Recognition; Production effect; Distinctiveness; Strength; Source memory.

There exist just a few simple mnemonic techniques that offer consistent and reliable memory enhancement. Well-established techniques include imagery (Paivio, 1971), elaboration (i.e., levels of processing; Craik & Lockhart, 1972; Craik & Tulving, 1975), generation (Slamecka & Graf, 1978; for a review and meta-analysis see Bertsch, Pesta, Wiscott, & McDaniel, 2007), and enactment (e.g., Cohen, 1981; Engelkamp & Krumnacker, 1980; for reviews see Engelkamp, 1998; Zimmer et al., 2001,). Each of these mnemonics offers an easily executed mechanism to quickly enhance memory, but the list of available techniques is remarkably short. We have been investigating another technique that may be worthy of being added to this list.

The phenomenon is very simple: When some material is read aloud and other material is read silently, memory is better for the material read aloud. First demonstrated by Hopkins and Edwards (1972), this effect has subsequently been rediscovered several times (Conway & Gathercole, 1987; Dodson & Schacter, 2001; Gathercole & Conway, 1988; MacDonald & MacLeod, 1998). Considering that it has been known to exist for nearly 40 years, and that it appears to rival in magnitude other well-known mnemonic techniques, this phenomenon has received surprisingly

Address correspondence to: Jason D. Ozubko, Rotman Research Institute, Baycrest Centre for Geriatric Care, 3560 Bathurst Street, Toronto, Ontario M6A 2E1, Canada. E-mail: jozubko@research.baycrest.org

This research was supported by Discovery Grant A7459 from the Natural Sciences and Engineering Research Council of Canada. We thank Brianna Groot and Niki Kowalski for their assistance in collecting the data in Experiment 3.

little attention until recently. In 2010 MacLeod, Gopie, Hourihan, Neary, and Ozubko offered a thorough investigation of this phenomenon, labelling it *the production effect*.

In their investigation MacLeod et al. (2010) demonstrated that the production effect was remarkably robust, having a measurable impact on memory even when words were first enhanced through generation or deep processing (see also Forrin, Jonker, & MacLeod, in press). Furthermore they demonstrated that production critically depended on a unique response being made to each word during study (i.e., saying "yes" or tapping a key did not lead to a production effect, whereas reading each word or even mouthing each word did; see also Forrin, Ozubko, & MacLeod, 2012). Taken together, these findings led MacLeod et al. to champion a distinctiveness account of the production effect (an idea originally offered by Gathercole & Conway, 1988), which held that distinctive processing of aloud items at study was critical to obtaining the effect (for more on distinctiveness, see Hunt, 2006Hunt, 2013; Hunt & Worthen, 2006).

Continuing research from our laboratory has confirmed that production is an effective and reliable manipulation across a wide variety of conditions. Production has been shown to increase memory in both recognition and recall for both younger and older individuals (Lin & MacLeod, 2012), to benefit both recollection and familiarity (Ozubko, Gopie, & MacLeod, 2012), to work in multiple forms (including mouthing, handwriting, and typing, Forrin et al., 2012), to be effective even if another individuals does the production (MacLeod, 2011), to protect against intentional forgetting (Hourihan & MacLeod, 2008), to have long-lasting benefits (Ozubko, Hourihan, & MacLeod, 2012), and to be applicable to educationally relevant materials (i.e., textbook chapters; Ozubko, Hourihan, et al., 2012).

Under a distinctiveness account of production, reading some words aloud enhances the distinctiveness of those items because it provides extra mnemonic information (i.e., a memory of having spoken those words) not present for silently read words. During a memory test, participants can strategically use this extra mnemonic information to infer which words were studied because realising that a word was recently spoken aloud confirms that it was recently studied. Consistent with this account, in mostly early studie, no production effect was observed when all words were spoken aloud (i.e., in between-participants conditions; Hopkins & Edwards, 1972; MacLeod et al., 2010; although see Gathercole & Conway, 1988). As well, in list-discrimination experiments when ''aloudness'' information was not diagnostic of list membership no production effect was obtained (Ozubko & MacLeod, 2010). However, a recent meta-analysis demonstrated that a between-participants production effect does appear to exist, albeit a small effect compared to the considerably more robust within-participants production effect, and is often non-significant within individual studies (Fawcett, 2013). Furthermore, recent work with the list-discrimination paradigm has demonstrated that this paradigm might not actually be able to support distinctiveness accounts of production given that attribution processes appear to play an important role (Bodner & Taikh, 2012). These new studies therefore raise the question of whether accounts other than the distinctiveness account might provide better explanations of the production effect.

In contrast to the distinctiveness account, one class of alternative explanations of the production effect could derive from strength-based accounts. By this type of account, reading words aloud strengthens the representations of those items more so than does reading words silently, and therefore words read aloud are easier to recognise and recall than are words read silently. Strength-based accounts would appear to be simpler than distinctiveness accounts inasmuch as there is no need to make any assumptions about differential contents in memory, and instead the theoretical focus can simply be on the relative strengths of memory representations.

Given the recent challenges to the distinctiveness account of the production effect, our goal in this article is to evaluate the viability of the alternative strength-based account of the production effect and to determine how it fares against the distinctiveness account. Specifically, we will examine predictions of strength-based accounts as they relate to the participant's ability not only to recognise whether a word was studied, but also to accurately judge the studied mode of words (i.e., ''aloud'', ''silent'', or ''new''). To understand how this approach will pit the strength-based account against the distinctiveness account, we must first consider how study mode judgements are hypothesised to occur under strength-based versus distinctiveness-based accounts.

JUDGING THE MODE OF STUDIED ITEMS

Ozubko, Gopie, et al. (2012) reported a production study in which participants were required not only to recognise words, but also to judge their studied mode (i.e., ‘‘aloud’’, ‘‘silent’’, or ‘‘new’’; see also Conway & Gathercole, 1987). Interestingly, participants were more accurate at identifying aloud words as ‘‘aloud’’ than they were at identifying silent words as ‘‘silent’’, as if it was easier to be sure that a word had been read aloud at study than that it had been read silently. Most important, however, participants did demonstrate study mode discrimination, and were most likely to call an aloud word ‘‘aloud’’ and to call a silent word ‘‘silent’’. At the time, Ozubko, Gopie, et al. (2012) compared the fits of their data with a single-process model, which was similar to and compatible with strength-based accounts in presuming that memory was subserved by a single dimension of strength, to those with a dual-process model, which was compatible with and similar to distinctiveness accounts in that it presumed that recognition is partly based on the recollection of the contextual details surrounding encoding, and the featural details of stimuli. It was reported that fits of single-process (i.e., strength-based) models were generally less optimal than fits of dual-process models that assumed that participants could use distinctive recollective details at test to infer which items had been studied (i.e., models compatible with the distinctiveness account). However, due to the limitations of the analysis techniques, these model tests omitted the study mode judgement experiment and instead analysed data from two experiments assessing the recollection and familiarity of words (seeYonelinas, 2002, for a review of recollection and familiarity). A dual-criterion strength-based account of study mode judgements was therefore not definitively ruled out.

If we assume that words read aloud are represented more strongly than are words read silently, then participants could presumably make reasonably accurate aloud/silent/new judgements based purely on memory strength if they responded ‘‘aloud’’ to the most strongly represented words, ‘‘silent’’ to more weakly represented words, and ‘‘new’’ to the weakest words. Supporting this proposal, participants do indeed believe memory to be better for words read aloud than for words read silently (Castel, Rhodes, & Friedman, 2013), so it seems likely that they would adopt this strategy if they based study mode judgements on memory strength alone. By a strength-based account, then, to perform study mode judgements in a production effect experiment, participants would need to adopt two decision criteria. If the strength of an item surpassed the higher criterion, participants could respond ‘‘aloud’’, if the strength of an item did not pass the higher criterion but did pass the lower criterion, participants could respond ‘‘silent’’, and for all other (weaker) items subjects could respond ‘‘new’’ (see Donaldson, 1996, for more on dual-criteria models). Under this model the probability of an ‘‘aloud’’ response should be directly related to the relative strength of the item in question. In contrast to the strength-based account, a distinctiveness account holds that, when deciding about study mode, participants consult the contents of memory for different types of information, rather than simply assessing the strength of representations. Specifically, participants query memory for evidence that a word has recently been read aloud. If that evidence is retrieved, then participants can confidently respond ‘‘aloud’’. If that evidence is not retrieved, participants respond ‘‘silent’’ if there was some evidence that the word was studied or, failing that, simply respond ‘‘new’’.

Both accounts can accommodate the types of study mode judgements observed in Ozubko, Gopie, et al. (2012) because both the strength of items and the contents of memory representations could be used to successfully differentiate study mode in a simple recognition design. The key to differentiating these accounts, then, is to experimentally manipulate the strength of representations such that the stronger representations are no longer those for the items that were spoken aloud. If participants are making study mode judgements based solely on the strength of representations, then increasing the strength of the silently read items should lead to a corresponding increase in ‘‘aloud’’ judgements for those items. Critically, if the strength of silently read items became equivalent to the strength of aloud words, it should not be possible to discriminate whether words had been read aloud or silently at study based solely on item strength. The only way to successfully judge study mode in this situation would be to consult the contents of their representations (i.e., ‘‘do I remember actually speaking this word aloud?’’). Demonstrating that participants can make study mode judgements independent of the relative strength of silently read words would therefore provide clear evidence in favour of the distinctiveness account over the strength-based account. This is what we set out to test.

It should be noted that the approach—to equate strength in two conditions or situations—has been successfully used previously in the memory literature to investigate strength-based accounts of other phenomena. For example, Hirst et al. (1986); Hirst, Johnson, Phelps, & Volpe, 1988) equated amnesic patients and controls on recognition so that they could examine their recall for differences. Considerably more relevant to our present study, Bink, Marsh, and Hicks (1999) equated recognition memory strength via repetition for viewed pictures vs imagined pictures at two retention intervals. Their goal was to examine source judgements without what essentially would otherwise amount to a strength confound between the conditions. Their evidence was consistent with the source information being present and used even when strength was equated. This logic is precisely the logic that underpins our study.

In Experiment 1 we increased the strength of silently studied words through repetition at study. Because, according to strength-based accounts, study mode judgements are based solely on strength, increasing the strength of silent items through repetition should lead to a corresponding increase in ‘‘aloud’’ ratings for those items at test. Consequently, we should observe a direct relation between the proportion of ‘‘aloud’’ ratings and memory strength, and the accuracy of study mode judgements for repeated silent items should decline in general.

In Experiment 2 we not only increased the strength of some silent items, we increased the strength of some silent items such that there was no significant difference between the overall recognition of repeated silent items and of aloud items. According to a strength-based account, as the strength of silent items becomes equivalent to that of aloud items, not only should the accuracy of study mode judgements continue to decline for those silent items, but at some point it should become impossible to differentiate aloud words from these strong-silent words. Importantly then, in Experiment 2 according to the strength-based account the ability to discriminate between aloud and silent words should break down. We should observe a similar proportion of ‘‘aloud’’ responses to both aloud and repeated silent items, as well as a similar proportion of ‘‘silent’’ responses to these two types of items.

In Experiment 3 we tested the strength-based account in a slightly different manner. In Experiments 1 and 2 words were studied aloud, twice-studied silent, or once-studied silent during study and participants were asked to make ‘‘aloud’’, ‘‘silent’’, and ‘‘new’’ ratings at test. In Experiment 3 participants were asked to make ‘‘aloud’’, ‘‘twice silent’’, ‘‘once silent’’, and ‘‘new’’ ratings during test. The rationale for Experiment 3 from the standpoint of the strength account is that if participants are relying on the strength of items to help determine study mode, then when aloud items are not identified as ‘‘aloud’’, participants should tend to identify these items as strong by selecting ‘‘twice silent’’. The distinctiveness account conversely predicts that when the study mode of an aloud item cannot be remembered, participants should be as likely to select ‘‘twice silent’’ for that item as they are for any other item where study mode cannot be remembered: In other words, when the ‘‘aloud’’ record is missing, assignment of mode should be based on guessing.

Finally we note that, although some researchers have investigated the accuracy of study mode judgements for aloud and silent words (Franck et al., 2000; Ozubko, Gopie, et al., 2012) and other researchers have examined whether reading aloud or silent helps text comprehension (Hale et al., 2011), no one has directly examined study mode judgements of aloud and silent words under strength manipulations. Specifically, examining how study mode judgements of aloud and silent words may contribute to the mnemonic benefit of reading aloud at study has never been investigated.

Across three experiments, then, our goal is to demonstrate that study mode knowledge is available independent of strength, and that study mode knowledge may therefore be contributing to memory performance in the production effect. If study mode judgements are affected by strength, we would then have good evidence for the strength-based account of production. If these ratings are not affected by strength, however, this would provide good evidence that distinctiveness accounts of production are more viable than strength-based accounts. The distinctiveness account does not predict that the proportion of ‘‘aloud’’ ratings should correlate with overall memory strength, and instead predicts that the ability to discriminate between aloud and silent words should remain strong in these experiments.

Consequently these experiments will serve as a critical test of the predictions of the strength-based versus the distinctiveness account of the production effect.

EXPERIMENT 1

The goal of Experiment 1 was to examine the study mode judgements of words read aloud versus silently at study. Critically, however, we included two sets of silently studied words, one engineered to be of greater memory strength than the other. Previous work has shown that when words are repeated at study, hit rates increase proportionally to the number of times an item was studied (e.g., Hintzman, Curran, & Oppy, 1992). Assuming that hit rates index memory strength, a straightforward method to increase the memory strength of silent items should be to present some of those items multiple times at study. Thus, in Experiment 1, participants studied three types of items: words read aloud, once-studied silent words, and twice-studied silent words. On the later item recognition test old words were mixed with new words and participants decided whether each word was studied aloud, or silently, or was new.

Although past studies have demonstrated that individuals can recall contextual information about studied items, information such as source (e.g., Johnson, Hashtroudi, & Lindsay, 1993) or background (Hockley, 2008; Smith, 1979) at the time of learning, the purpose of Experiment 1 was specifically to evaluate whether, for the production effect, ratings of aloud/silent status scaled with memory strength. Thus we were less interested in whether individuals can successfully identify study mode and more interested in whether there is evidence that these ratings are affected by memory strength. Once again, the strength-based account predicts that ''aloud'' ratings should scale linearly with memory strength, and so should be significantly greater for twice-studied than once-studied words. The distinctiveness account makes no such prediction; instead it expects more ''aloud'' ratings for words read aloud at study compared to words read silently at study, but no difference in ''aloud'' ratings for once-studied or twice-studied silent words. This would be the case because the ''aloud'' information would be available in memory only for the items actually studied aloud.

Method

Participants. A total of 26 Wilfrid Laurier University students participated in exchange for course credit.

Stimuli. A pool of 120 words was obtained from MacDonald and MacLeod (1998, Appendix A). These words ranged from 5 to 10 letters long with frequencies greater than 30 per million (Thorndike & Lorge, 1944). In all phases of the experiment words were presented in lowercase 18-point Courier font at the centre of the screen.

Procedure. Participants were told that they were participating in a memory experiment consisting of a study phase followed by a test phase. They were instructed that words would appear on the computer screen individually and would progress automatically. Words were presented for 2000 ms each with a 500-ms inter-stimulus interval. Participants were instructed that words at study would appear in either blue or white font, and that they were to read blue words aloud and white words silently (without moving their lips).

At study, participants saw 20 blue words, 20 white words presented once (once-studied), and 20 white words presented twice (twice-studied). Thus participants saw 60 words during the study phase, but 80 trials because 20 of those words repeated. Immediately following the study phase participants were presented with 120 individual words and asked to determine whether each word had been read aloud during study (''aloud''), had been read silently during study (''silent''), or had not been presented during study (''new''). Participants pressed ''m'' to indicate ''aloud'', ''spacebar'' to indicate ''silent'', and ''c'' to indicate ''new''. The 120 test items consisted of the 60 studied words randomly intermixed with 60 new words, and all words were presented in yellow to avoid colour associations from the study phase.

Results and discussion

To analyse the performance of aloud, twice-studied silent, once-studied silent, and new items, we began by collapsing ''aloud'' and ''silent'' responses into ''old'' responses, so that hit and false alarm rates could be calculated. Mean false alarm rate was .33 with a standard error of .04. Initial analyses of the hit rates revealed that hit rates were significantly higher for aloud items

($M = .88$, $SE = .02$) than for twice-studied silent items ($M = .81$, $SE = .03$) or for once-studied silent items ($M = .67$, $SE = .04$), $t(25) = 2.55$, $p < .05$, $d = 0.60$, and $t(25) = 6.20$, $p < .01$, $d = 1.38$, respectively. Hit rates for twice-studied silent items were also significantly higher than hit rates for once-studied silent items, $t(25) = 3.61$, $p < .01$, $d = 0.80$. Using hit rates to index overall memory strength, then, strength was greatest for words read aloud at study, intermediate for twice-studied silent items, and weakest for once-studied silent items. Clearly there was a strong production effect, sufficiently strong that a single production even led to better performance than silent repetition.

The critical question now is whether "aloud" ratings scale with memory strength (as the strength-based account predicts). To test this prediction we first analysed the raw study mode judgements. Interpretation of these raw judgements should be taken with caution, as the relation between item recognition, source memory, and bias are naturally inter-mixed in the overall raw scores. Nonetheless we believe that an analysis of the raw scores can be informative. Following that analysis, the data from Experiment 1 will be fit using the Batchelder-Riefer model of source monitoring, a commonly used multinomial model of source memory (see Batchelder & Riefer, 1990, 1999; Bayen, Murnane, & Erdfelder, 1996; Dodson, Prinzmetal, & Shimamura, 1998). The benefit of this analysis is that it will separate and quantify item recognition, source memory, and bias as separate parameter estimates. Analysis of these parameters will therefore provide a more systematic evaluation of the conclusions that we are able to draw from our evaluation of the raw judgements scores.

Raw study mode judgements for Experiment 1 are presented in Figure 1A. In terms of source accuracy, on first glance it appears as if silent items were more accurately identified than aloud items. That is, for twice-studied silent and once-studied silent items, there were more "silent" than "aloud" responses, $t(25) = 10.33$, $p < .01$, $d = 3.56$ and $t(25) = 11.12$, $p < .01$, $d = 3.34$, respectively. However, for aloud items, there were equivalent "aloud" and "silent" responses, $t(25) = 0.71$, $p = .48$. And yet, a strong bias exists in these data: there is an overwhelming tendency being to identify all items as "silent". This is exemplified best in the responses to new items, where participants responded "silent" considerably more frequently than "aloud" even though these items were never studied, $t(25) = 7.98$, $p < .01$, $d = 2.04$. Similar biases in the context of production have been reported by Dodson and Schacter (2001), Ozubko, Gopie, et al. (2012), and Bodner and Taikh (2012). It has been argued (Dodson & Schacter, 2001) that this bias arises because participants adopt stricter decision criteria for the more memorable stimulus class (i.e., the aloud items) but, regardless of why the bias occurs, its presence means that the raw study mode judgements can be misleading.

The observation that new, unstudied items are frequently being identified as "silent" raises the question of whether the frequent tendency to identify twice-studied and once-studied silent items as "silent" is being driven by bias rather than by source memory. Consequently the most accurate way to gauge study mode judgement accuracy from these raw data is to evaluate how much more likely participants are to judge the correct source for an item as compared to how likely they are to misjudge any other item as coming from that source. Examining the data in this way, aloud items were more likely to be rated "aloud" than were items from any other category (twice-silent, once-silent, and new), $F(1, 25) = 70.16$, $MSe = 0.03$, $p < .01$, $p\eta^2 = .74$. Similarly, twice-studied silent and once-studied silent were more likely to be rated "silent" than were aloud items and new items, $F(1, 25) = 40.53$, $MSe = 0.03$, $p < .01$, $p\eta^2 = .62$, and $F(1, 25) = 28.01$, $MSe = 0.02$, $p < .01$, $p\eta^2 = .53$, respectively. Hence, participants could accurately identify the study modes of aloud and silent items, despite the fact that there was a strong bias to identify all items as "silent".

The only evidence that study mode ratings may be scaling with memory strength is a small but significant increase in the proportion of "aloud" responses to twice-studied silent items compared to once-studied silent items, $t(25) = 2.10$, $p < .05$, $d = 0.35$. However, this likely is just an artefact of more twice-studied silent items being recognised compared to once-studied silent items, as "silent" responses also increased for twice-studied silent items compared to once-studied silent items, $t(25) = 2.13$, $p < .05$, $d = 0.54$. To provide a more sophisticated evaluation of these data, we turn now to our modelling efforts.

The Batchelder-Riefer model of source monitoring is a multinomial approach to modelling source judgements that separates item recognition, source memory, and bias (Batchelder & Riefer, 1990, 1999; Bayen et al., 1996; Dodson et al., 1998). Specifically, the model assumes that

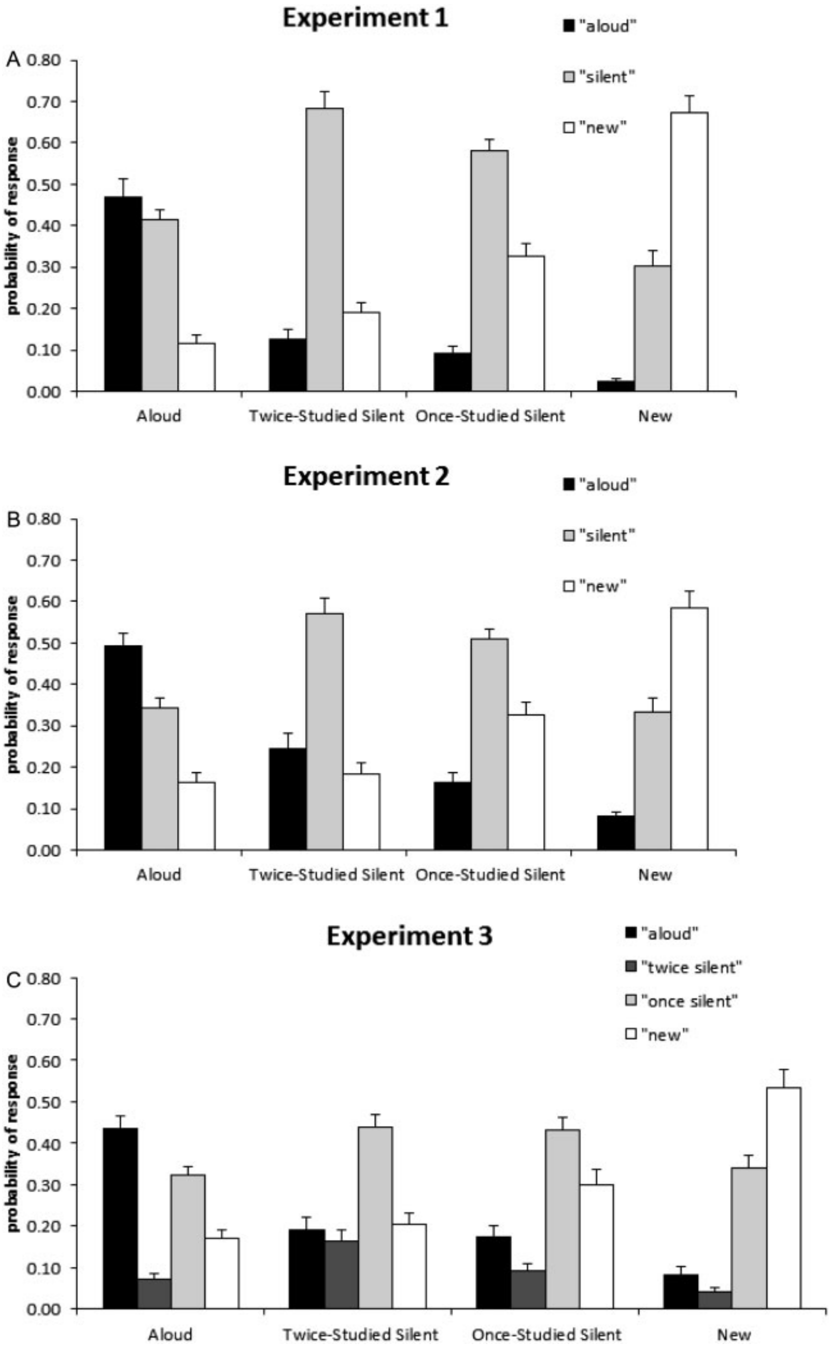

Figure 1. Mean proportions of study mode judgements in Experiments 1 through 3. Error bars are standard errors of the corresponding means.

a participant's response to any given item during the test is a result of a hierarchical series of probabilistic decisions. To provide an overview, the model assumes that on any given trial participants recognise an item as "old" with some probability (D_i; item recognition). For recognised items there is a probability that the source is known and can be immediately identified (d_i; source memory). For items where the source is not known, participants will guess between the sources, and the model represents the likelihood of these guesses using another parameter (*a*; bias given item was recognised). For items that are not recognised and for new items, participants sometimes will identify the item as "new", but other times will guess "old" (*b*; guessing bias).

For items that are guessed to be old, source will once again be guessed and once again represented using another parameter (g; bias given a guess). Using the equations provided by Dodson et al. (1998) as a guide, we implemented a multinomial model, slightly modified to account for three encoding conditions (aloud, twice silent, once silent) but only two response categories ("aloud" and "silent"). The model was implemented in Excel and we used the Excel solver to fit this model to the data from each participant in Experiment 1 individually.

For each participant we obtained estimates of item recognition, source memory, and bias using the Batchelder-Riefer multinomial model. One participant's data could not be fitted to the model but all other participants fitted well. These data are presented in the first column of Table 1. In terms of item recognition the D_i parameters estimate the probability of identifying a studied item as "old" separate from bias and source memory. These results parallel our overall hit rate analysis. Specifically, aloud items were more recognisable than twice-studied silent and once-studied silent items, $t(25) = 2.00$, $p < .05$, $d = 0.64$, and $t(25) = 6.49$, $p < .01$, $d = 1.80$, respectively, and twice-studied silent items were more recognisable than once-studied silent items, $t(25) = 2.93$, $p < .01$, $d = 0.92$.

In terms of source identification the d_i parameters represent the probability of correctly remembering the source (i.e., study mode) of an item, given that the item was recognised. It is important to note that this does not represent the overall likelihood that a participant will correctly identify the study mode of an item, just the likelihood that the source will be remembered if the item is recognised. Hence there is no expectation

TABLE 1
Experiments 1, 2, and 3

Parameters		*Experiment 1*	*Experiment 2*	*Experiment 3*
p(Item Recognition) (D_i)				
	Aloud	.84 (.03)	.72 (.03)	.66 (.04)
	2 × Silent	.73 (.04)	.70 (.03)	.61 (.05)
	Silent	.56 (.03)	.46 (.04)	.43 (.05)
p(Source Identification) (d_i)				
	Aloud	.38 (.05)	.36 (.05)	.43 (.06)
	2 × Silent	.37 (.08)	.27 (.06)	.14 (.04)
	Silent	.38 (.05)	.36 (.05)	.14 (.05)
Bias given item was recognised as old but source was not known (a)				
	p("Aloud")	.27 (.04)	.42 (.04)	.31 (.05)
	p("2 × Silent")	–	–	.15 (.02)
	p("Silent")	.73 (.04)	.58 (.04)	.54 (.04)
Bias given item was guessed to be "old" (g)				
	p("Aloud")	.11 (.02)	.20 (.02)	.16 (.03)
	p("2 × Silent")	–	–	.08 (.02)
	p("Silent")	.89 (.02)	.80 (.02)	.76 (.03)
p(Guess "Old") (b)		.31 (.04)	.42 (.04)	.46 (.04)

Means (and standard errors) of parameters estimated from the Batchelder-Riefer multinomial model used to fit the data from Experiments 1 through 3.

that these values should be .50 by chance. In terms of the probability of correctly remembering the study mode, then, as we concluded from our analyses of the raw data, there was no significant difference among the item types (aloud, twice-studied silent, or once-studied silent), $F(2, 48) < 1$. Finally participants were biased to respond ‘‘silent’’ in the absence of memory for study mode, both when items could be correctly recognised at the item level (a parameter), $t(24) = 4.97$, $p < .01$, $d = 2.09$, and when items could not be recognised at the item level but were guessed to be old (g parameter), $t(24) = 18.47$, $p < .01$, $d = 7.80$.

In sum, upon initial examination of the raw data it appeared as if silent items could be more easily identified than aloud items, but further analyses reveal that there is no difference in source discrimination between aloud and silent items and that any indication otherwise is an artefact of bias. This interpretation was supported by an evaluation of the raw study mode judgements and by a multinomial modelling approach. More important, however, there is no evidence that ‘‘aloud’’ ratings necessarily scale with memory strength. Both our evaluation of the raw scores and the model's parameter estimates yield no evidence that twice-studied silent items are more likely to be identified as ‘‘aloud’’ than are once-studied silent items.

EXPERIMENT 2

The goal of Experiment 2 was to compare study mode judgements of aloud and silently studied words when the strength of some silent words was more comparable to that of words read aloud at study. According to a strength-based account the proportion of ‘‘aloud’’ responses should increase in proportion to the memory strength of a stimulus class. Importantly, however, if two sets of stimuli have comparable memory strengths, it should not be possible to discriminate which class was spoken aloud and which class was read silently.

One key concept in terms of memorability is that the more unique a stimulus is, the more memorable it is (see Hunt, 2006). We leveraged this relation between uniqueness and memorability in an attempt to keep the methods of Experiments 1 and 2 as similar as possible, while still making the memorability of aloud words and twice-studied silent words more equivalent. In Experiment 2 we therefore increased the number of words to be read aloud at study, while keeping the number of once-studied and twice-studied silent items the same as in Experiment 1. In Experiment 1 only 33% of study words were aloud, with 25% of study trials aloud. These aloud trials could therefore have benefited from being quite unique, making them more memorable than twice-studied silent words. By equating the number of aloud and silent trials in Experiment 2—60% aloud (20% each once-studied silent and twice-studied silent), resulting in 60 aloud trials and 60 silent trials—the memorability of aloud trials in general would be expected to decline, ideally becoming equivalent to that of twice-studied silent items.

Equating the number of aloud and silent trials in Experiment 2 also served several methodological purposes. First, because aloud trials were less common than silent trials in Experiment 1, the enhanced memorability for aloud items in general (i.e., the production effect) could have been due to this uniqueness. Given the numerous findings suggesting that the production effect arises when an equal number of aloud and silent words are studied, and so is not dependent on being a rare trial type, this possibility is remote. Nonetheless, it should be addressed. Furthermore, the fact that 75% of the study list in Experiment 1 was silently studied may explain why participants were biased to respond ‘‘silent’’ so frequently in Experiment 1: In the absence of any evidence, ‘‘silent’’ was the best guess. Experiment 2 therefore should serve both to render aloud and twice-studied silent words more similarly memorable and to equate the number of aloud and silent trials during the study phase to address any methodological issues that might have been present in Experiment 1.

Method

Participants. A total of 28 students participated in exchange for course credit; 18 were from Wilfrid Laurier University and 10 were from the University of Waterloo.

Stimuli. A larger word pool was required for this experiment, so a pool of 384 nouns was downloaded from the MRC Linguistic Database (http://websites.psychology.uwa.edu.au/school/MRCDatabase/uwa_mrc.htm). All words were between 5 and 10 letters long, were within one standard deviation of the mean on the measures of concreteness and imageability used by the database, and had frequencies greater than 30 per million, making them similar to the pool used in Experiment 1.

Procedure. The procedure for this experiment was identical to that of Experiment 1, except that the number of blue (aloud) words presented at study was increased to 60. Thus, at study, participants saw 120 items: 60 blue words, 20 white words presented once, and 20 white words presented twice. At test, the 100 studied words were mixed with 100 new words and participants decided "aloud", "silent", or "new" for each test word.

Results and discussion

As in Experiment 1, aloud and silent responses were first collapsed to generate hit and false alarm rates. The mean false alarm rate was .38, with a standard error of .06. In terms of hit rates, this time there was no difference between the hit rates for aloud items ($M = .81$, $SE = .03$) and twice-studied silent items ($M = .84$, $SE = .03$), $t(17) = 1.05$, $p = .31$. Hits rates for both aloud items and twice-studied silent items were greater than for once-studied silent items ($M = .68$, $SE = .04$), $t(17) = 4.23$, $p < .01$, $d = 0.80$, and $t(17) = 5.70$, $p < .01$, $d = 1.00$, respectively. Thus, in Experiment 2, hit rate performance was statistically equated for aloud and twice-studied silent items. Using hit rates as an index of strength, we will assume that the strength of aloud items and twice-studied silent items was roughly equivalent.

Raw study mode judgements for test probes are shown in Figure 1B. An across-experiment ANOVA revealed that there was no overall difference between the study mode judgements of Experiment 1 and Experiment 2, $F(1, 42) < 1$. A study mode judgement ("aloud" vs "silent") by experiment interaction was observed, $F(1, 42) = 131.82$, $MSe = 0.05$, $p < .05$, $p\eta^2 = .10$, driven by the fact that there were small variations in the "aloud" and "silent" study mode judgements in Experiment 2 compared to Experiment 1. No other interactions that included experiment as a factor were significant, all Fs < 1. Given that there are no striking differences between the ratings of Experiment 1 and 2, this minimises methodological concerns surrounding the fact that in Experiment 1 only 25% of the study trials were aloud items. Having an unequal number of aloud and silent trials in Experiment 1 clearly was not the source of the "silent" bias in study mode judgements, and the enhanced memory for the aloud items over silent items cannot be attributed solely to having fewer aloud trials at study in Experiment 1.

As was the case in Experiment 1, raw study mode judgements were analysed here with some caution because item recognition, source memory, and bias are naturally intermixed in the overall raw scores. In terms of source accuracy there were more "silent" than "aloud" responses to twice-studied silent and once-studied silent items, $t(27) = 4.43$, $p < .01$, $d = 1.56$, and $t(27) = 8.93$, $p < .01$, d $= 2.68$, respectively, and there were more "aloud" than "silent" responses to aloud items, $t(27) = 2.99$, $p < .01$, $d = 1.02$. As in Experiment 1, straightforward interpretation of these effects is not possible because a strong bias exists in the data. Participants generally responded "silent" to all items, best exemplified in the tendency to respond "silent" more than "aloud" to new items, $t(27) = 7.13$, $p < .01$, $d = 1.77$. Importantly, the presence of this bias in Experiment 2, where there were equal numbers of silent and aloud items during study, demonstrates that the bias to respond "silent" does not depend on the proportion of silent items presented at test.

In light of this bias, once again the most accurate way to gauge source accuracy from these raw data is to evaluate how much more likely participants were to judge the correct source of an item as compared to how likely they were to misjudge any other item as coming from that source. Examining the data in this way, aloud items were more likely to be rated "aloud" than any other category (twice-silent, once-silent, and new) was to be rated "aloud", $F(1, 27) = 87.88$, $MSe = 0.02$, $p < .01$, $p\eta^2 = .77$. Similarly, twice-studied silent and once-studied silent were more likely to be rated "silent" than aloud and new items were to be rated "silent", $F(1, 27) = 28.35$, $MSe = 0.03$, $p < .01$, $p\eta^2 = .51$, and $F(1, 27) = 34.32$, $MSe = 0.01$, $p < .01$, $p\eta^2 = .56$, respectively. Hence participants could accurately identify the study mode of aloud and silent items, despite the fact that there was a strong bias to identify all items as "silent".

Like Experiment 1, then, the only evidence from the raw judgement scores consistent with the idea that study mode ratings might be scaling with memory strength was a small but significant increase in the proportion of "aloud" responses to twice-studied silent items compared to once-studied silent items, $t(27) = 3.15$, $p < .01$, $d = 0.55$. However, this once again is likely an artefact of the fact that more twice-studied silent items were recognised compared to once-studied silent items. ‘‘Silent’’ responses marginally increased for twice-studied silent items compared to once-studied silent items, $t(27) = 1.49$, $p = .07$, $d = 0.34$. To provide a more sophisticated evaluation of these data, we turn now to our modelling efforts.

For each participant we obtained estimates of item recognition, source memory, and bias using the same Batchelder-Riefer multinomial model as in Experiment 1. Again, one participant's data could not be fitted to the model but all other participants fitted well. These data are presented in the second column of Table 1. In terms of item recognition (D_i), these results parallel our overall hit rate analysis—aloud items and twice-studied silent were more recognisable than once-studied silent items, $t(26) = 6.38$, $p < .01$, $d = 1.52$, and $t(26) = 5.97$, $p < .01$, $d = 1.30$, respectively, but did not differ from each other, $t(26) = 0.23$, $p = .59$. In terms of source identification (d_i) there was no significant difference among the three item types (aloud, twice-studied silent, or once-studied silent), $F(2, 52) = 2.12$, $MSe = 0.04$, $p = .13$. Individual comparisons of twice-studied silent vs aloud and twice-studied silent vs once-studied silent confirmed the absence of any difference, $t(26) = 1.46$, $p = .16$, and $t(26) = 1.45$, $p = .16$, respectively. Finally, participants were biased to respond ‘‘silent’’ in the absence of memory for study mode. This effect was marginal when items could be correctly recognised at the item level (a parameter), $t(26) = 1.38$, $p = .09$, $d = 0.70$, but significant when items could not be recognised at the item level but were guessed to be old (g parameter), $t(26) = 13.72$, $p < .01$, $d = 5.45$. Thus, there was actually some reduction in ‘‘silent’’ bias in Experiment 2, at least when words could be recognised at the item level.

In sum, the results of Experiment 2 closely resemble those of Experiment 1. We found no evidence that ‘‘aloud’’ responses scale with memory strength and importantly, although the memory strength of aloud and twice-silent study items was roughly equated in Experiment 2, participants could still identify the study mode of items. According to a strong strength-based account this should not have been possible, and according to a more moderate strength-based account some evidence that ‘‘aloud’’ ratings scale with strength should have been observed.

EXPERIMENT 3

Across two experiments we have examined source memory for study mode in the production effect. No evidence has been observed yet to indicate that participants use the strength of an item's study modality to make source memory judgements. So far, however, both Experiments 1 and 2 have used the strategy of examining whether strong silent items (i.e., twice-studied silent items) are more likely to be mistaken for aloud items at test. To provide convergence, Experiment 3 uses a slightly different approach.

In Experiment 3 at test participants are asked to identify each test probe as either ‘‘aloud’’, ‘‘twice silent’’, ‘‘once silent’’, or ‘‘new’’. The purpose of including two silent response categories is twofold. First, it will allow us to investigate the possibility that strength is playing a role in study mode discrimination insomuch as aloud items, when they are not identified as ‘‘aloud’’, will be likely to be identified as ‘‘twice silent’’ (i.e., the stronger silent condition). In contrast to this strength-based prediction, the distinctiveness account would not expect aloud items to be identified as ‘‘twice silent’’ more so than any other items (once-studied silent or new) would be misidentified as ‘‘twice silent’’.

A second issue that Experiment 3 will address is whether remembering that an item was said aloud is more distinctive than remembering that an item was read silently. That is, a more peripheral prediction of the distinctiveness account that we have not discussed in much detail thus far is that saying a word aloud should usually be more distinctive than is reading a word silently. This aspect of the distinctiveness account explains why at test, if a participant remembers ‘‘I said it aloud’’, it can help them identify a word as studied, whereas ‘‘I don't remember saying it aloud’’ cannot. When Ozubko, Gopie, et al. (2012) examined aloud/silent/new ratings in a typical production effect study, consistent with the distinctiveness account, there was indeed an advantage for aloud items in corrected recognition scores insomuch as participants were more likely to identify aloud words as ‘‘aloud’’ than they were to identify silent words as ‘‘silent’’.

In Experiments 1 and 2 participants were equally able to correctly remember the source of aloud, twice-studied silent, and once-studied silent items. The results of these two experiments would, then, seem at odds with the findings of

Ozubko, Gopie, et al. (2012) and with the distinctiveness account. However, the reason for this might simply have been that in Experiments 1 and 2 "silent" responses captured two sources of memory (twice-studied silent and once-studied silent) whereas "aloud" responses captured only one source of memory. The accuracy of "silent" ratings in Experiments 1 and 2 might have therefore been inflated due to the fact that participants did not have to specify the source of silent items precisely—so long as a participant did not recall a word had been spoken aloud, "silent" was a reasonable guess. In Experiment 3, because participants will be required to differentiate aloud, twice-studied silent, once-studied silent, and new, the distinctiveness account predicts that participants should show better source discrimination for aloud items than for twice-studied or once-studied silent items.

Method

Participants. A total of 24 students from the University of Waterloo participated in exchange for course credit.

Stimuli. Experiment 3 used the same stimuli as Experiment 2.

Procedure. The procedure for Experiment 3 was identical to that of Experiment 2 except that at test participants were asked to respond "aloud", "twice silent", "once silent", or "new" for each test word.

Results and discussion

As in Experiments 1 and 2, aloud and silent responses were first collapsed to generate hit and false alarm rates. The mean false alarm rate was .47, with a standard error of .04. In terms of hit rates, although numerically very similar, there was a significant difference between the hit rates for aloud items ($M = .83$, $SE = .02$) and twice-studied silent items ($M = .80$, $SE = .03$), $t(23) = 2.20$, $p < .04$, $d =$ 0.28. Hits rates for both aloud items and twice-studied silent items were greater than for once-studied silent items ($M = .70$, $SE = .04$), $t(23) = 4.72$, $p < .01$, $d = 0.90$, and $t(23) = 3.68$, $p < .01$, $d = 0.61$, respectively. Thus in Experiment 3 although hit rate performance was statistically not quite equated for aloud and twice-studied silent items, it was close and in the same range of difference, numerically, as Experiment 2. However, the goal of Experiment 3 was not necessarily to fully equate aloud and twice-studied silent items (this task was achieved in Experiment 2). Thus, so long as twice-studied silent items were stronger than once-studied silent items, the predictions of Experiment 3 are still testable. Therefore we can consider aloud and twice-studied silent items simply to be very close in strength and proceed with our analyses.

Raw study mode judgements for test probes are shown in Figure 1C. As in Experiments 1 and 2, raw study mode judgements were analysed here with some caution because item recognition, source memory, and bias are naturally intermixed in the overall raw scores. Also as in Experiments 1 and 2, a strong bias exists in the data. Participants generally responded "once silent" to all items, and this is exemplified best in the tendency to respond "once silent" more than "aloud" or "twice silent" to new items, $t(23) = 7.84$, $p < .01$, $d = 2.10$, and $t(23) = 10.05$, $p < .01$, $d = 2.68$, respectively. In light of this bias, once again the most accurate way to gauge source accuracy from these raw data is to evaluate how much more likely participants are to judge an item as the correct source compared to how likely they are to misjudge any other item as coming from that source. Examining the data in this way, aloud items were more likely to be rated "aloud" than any other category (twice-silent, once-silent, and new) was to be rated "aloud", $F(1, 23) = 74.72$, $MSe = 0.01$, $p < .01$, $p\eta^2 = .77$. Similarly, twice-studied silent items were more likely to be rated "twice silent" than any other category (aloud, once-silent, and new) was to be rated "twice silent", $F(1, 23) = 18.18$, $MSe = 0.01$, $p < .01$, $p\eta^2 = .44$, and once-studied silent items were more likely to be rated "once silent" than any other category (aloud, twice-silent, and new) was to be rated "once silent," $F(1, 23) = 7.43$, $MSe = 0.01$, $p < .05$, $p\eta^2 = .24$. Hence, participants could accurately identify the study mode of items, despite the fact that there was a strong bias to identify all items as "once silent".

Finally, aloud and once-studied silent items were more likely to be identified as "twice silent" than were new items, $t(23) = 2.11$, $p < .05$, $d = 0.54$, and $t(23) = 3.21$, $p < .01$, $d = 0.82$, respectively. Importantly, however, aloud items were no more likely to be identified as "twice silent" than were once-studied silent items, $t(23) = 1.42$, $p = .09$,[1]

[1] Note that this difference goes in the direction opposite to that expected under the strength account (i.e., p["twice silent"|once-studied silent] > p["twice silent"|aloud]).

indicating that ‘‘twice silent’’ ratings did not scale with memory strength. As well, twice-studied silent items were no more likely to be identified as ‘‘aloud’’ than were once-studied silent items, $t(23) = 0.08$, $p = .53$, indicating that ‘‘aloud’’ ratings also did not scale with memory strength.

For each participant we obtained estimates of item recognition, source memory, and bias using a version of the Batchelder-Riefer multinomial model that accommodated three sources of memory. These data are presented in the third column of Table 1. In terms of item recognition (D_i), these results parallel our overall hit rate analysis—aloud items were marginally more recognisable than twice-studied silent and significantly more recognisable than once-studied silent items, $t(23) = 1.51$, $p = .07$, $d = 0.24$, and $t(23) = 4.17$, $p < .01$, $d = 1.02$, respectively, and twice-studied silent items were more recognisable than once-studied silent items, $t(23) = 3.04$, $p < .01$, $d = 0.73$. In terms of source identification (d_i), source was remembered better for aloud items than for either twice-studied or once-studied silent items, $t(23) = 4.45$, $p < .01$, $d = 1.21$, and t(23) $= 3.76$, $p < .01$, $d = 1.13$, respectively. Source identification did not differ between twice-studied and once-studied silent items, $t(23) = 0.01$, $p = .99$. Finally, participants were biased to respond ‘‘once silent’’ in the absence of memory for study mode, both when items could be correctly recognised at the item level (a parameter), $F(1, 23) = 21.28$, $MSe = 0.05$, $p < .01$, $p\eta^2 = .48$, and when items could not be recognised at the item level but were guessed to be old (g parameter), $F(1, 23) = 196.34$, $MSe = 0.03$, $p < .01$, $p\eta^2 = .90$.

In sum, just as in Experiments 1 and 2, we found no evidence that ‘‘aloud’’ responses scaled with memory strength. As well, ‘‘twice silent’’ responses did not scale with memory strength: Aloud items and once-studied silent items were equally likely to be misidentified as ‘‘twice silent’’. Finally, consistent with the distinctiveness account, source identification accuracy was superior for aloud items than for either twice-studied or once-studied silent items.

GENERAL DISCUSSION

A strength-based account of the production effect would hold that memory strength is enhanced for items that are produced at study, and that it is this enhanced strength that underlies the later memory advantage. Under this view, if participants were asked to make study mode judgements (‘‘Was this item studied aloud, silently, or not at all?’’), they would likely infer (to some degree) that the strongest items were studied aloud, whereas intermediate and weaker items were read silently at study or were not studied. In contrast to this view, the distinctiveness account proposes that the act of production is encoded into memory and that at test participants can strategically access this information to make an inference about whether an item was studied (i.e., if you can remember speaking a word in the experiment, it is likely that you studied that word).

The finding that individuals can identify contextual information about studied items is not new, as the source memory literature certainly illustrates (Bink et al., 1999; for review, see Johnson et al., 1993). Indeed, study mode judgements of words read aloud or silently has also been examined to some degree in the past (Franck et al., 2000; Ozubko, Gopie, et al., 2012). However, with respect to the production effect, there are two principal new findings reported in this article. First, information about study mode clearly is available independent of memory strength, and second, information about having spoken a word aloud is more distinctive than information about having read an item silently.

Across three experiments we have demonstrated that participants are able to make accurate study mode judgements, and that they can do so independent of memory strength. Increasing the strength of items does not lead to the substantial increase in the proportion of ‘‘aloud’’ responses (Experiment 1). As well, roughly equating the memory strength of aloud and silently studied items does not undermine participants’ ability to make accurate study mode judgements (Experiment 2). And finally, when participants must differentiate ‘‘twice silent’’ and ‘‘once silent’’ during test, study mode memory is actually superior for aloud items compared to either of the silent conditions (Experiment 3). Overall, then, these results are generally inconsistent with a strength-based account of production, and instead support the distinctiveness explanation: At test, participants do have access to mnemonic information about which words they spoke aloud or read silently, independent of overall memory strength. Mode of encoding is preserved in memory, and ‘‘aloud’’ information is particularly distinctive, at least compared to ‘‘silent’’ information.

A critic might suggest that having asked participants to make study mode judgements instead of old/new judgements at test might have led them to adopt a recognition strategy in the current experiments that differs from the strategy commonly adopted in production effect recognition studies. In essence, perhaps in typical old/new recognition studies, strength differences do underlie the production effect, but in a source discrimination task they do not. Although we cannot be certain, we believe this possibility to be remote. In all three experiments we found no evidence that strength influences recognition. If strength did play a role during old/new recognition, then even if our modality judgement recognition test biased participants to adopt an alternative strategy, this tendency should not have been absolute. That is, some participants should have continued to use strength, at least to some degree. As a result we should have at least seen trends in the data that suggested a role for strength in recognition. We did not.

It should be noted that, although the impetus for this work was to address recent challenges to the distinctiveness account of production, both those challenges and this current work are consistent with one another, and with the larger literature regarding the production effect. That is, recent work by Bodner and Taikh (2012) has demonstrated that evidence from list-discrimination tasks, previously used to argue against strength-based accounts of production (Ozubko & MacLeod, 2010) may not be able to support such claims. However, Bodner and Taikh's work did not specifically endorse or provide evidence in favour of strength-based accounts. Instead they highlighted the limitation of the list-discrimination task for discriminating between accounts of the production effect, thereby reopening the question of strength-based accounts.

In a similar vein, using meta-analysis, Fawcett (2013) recently demonstrated that there is a small between-participants production effect across experiments. The strength-based account specifically predicts that there should be a between-participants production effect; however, a distinctiveness account could also allow for this possibility. If participants did study an entire list aloud and realised that recalling "aloudness" information at test would be helpful at discriminating old from new items, there is no reason why a between-participants production effect could not arise. The between-participants production effect may be less reliable than the within-participants production effect because, when participants do read all words in a study list aloud, the act of reading things aloud no longer seems distinct, and they might not consider using that information at test, or they might rely on it considerably less.

Although we argue that the new data reported here are incompatible with a strength-based account, does this mean that strength plays no role in the production effect? As we have discussed before (Ozubko, Gopie, et al., 2012; see also MacDonald & MacLeod, 1998), although we take the existing evidence to demonstrate that the distinctiveness account offers the best explanation of the main factor underlying the production effect, there may be some kind of strength-based component. Whether this component arises from enhanced attention or rehearsal of items spoken aloud, or some other less-intentional strategy, it is not unreasonable to suppose that a component of production could be strength-based. Indeed, we did observe a small rise in "aloud" responses to twice-studied silent items relative to once-studied silent items. Although we argued that this rise merely reflected an overall increase in recognisability of twice-studied silent items, it might be possible that this was a small strength-based effect. Indeed, although statistically the accuracy of study mode judgements did not differ between once-studied and twice-studied silent items in Experiment 2, numerically the accuracy of judgements for twice-studied silent items did decline, which could be indicative of a small strength-based influence on these judgements. In a similar manner, although we have suggested that the between-participants production effect (Fawcett, 2013) is compatible with the distinctiveness account, it could also be a reflection of the weaker, strength-based processes of production (MacLeod et al., 2010; see also Bodner & Taikh, 2012). Finally, production has recently been shown to rely at least in part on recollective processes, which are consistent with the use of distinctiveness, however production also has a familiarity-based component, which is compatible with a strength-based account (Ozubko, Gopie, et al., 2012). In sum, although the evidence in the literature suggests that the production effect relies mainly on distinctiveness, there may be a small influence of strength as well.

The implication that there may be two influences on production, even if one is dominant, raises interesting questions regarding whether these two influences could be selectively manipulated or could have different consequences. For example, the recent demonstration that a small between-participants production effect may exist (Fawcett, 2013) raises the possibility that this between-participants production effect may be solely strength-based. Certainly, a strength-based account would predict a between-participants production effect, and its smaller size is consistent with the notion of a smaller strength-based component of production. So, perhaps the between-participants production effect fundamentally differs from the within-participants production effect not just by being smaller, but by relying primarily on fundamentally different processing. Another interesting possibility is that these two influences of production may have different consequences. For example, strength may be applicable in recognition tests, where studied words are provided to participants and global familiarity can be used to judge old/new status, however strength may play no role in recall tests where studied words must be actively generated and retrieved by the participant. Strength and distinctiveness may also differ in their resistance to long-term forgetting, or in their ability to inform memory for more complex materials, such as educationally relevant texts (see Ozubko, Hourihan, et al., 2012).

The primary conclusion from the current findings is that, even if strength does play a small role in the production effect, distinctiveness appears to be the dominant factor contributing to this phenomenon, at least in within-participants designs. In sum, despite recent challenges to the distinctiveness account of production, the current data provide clear evidence consistent with the claim that distinctiveness does underlie this phenomenon. Regardless of the strength of items, participants are readily able to identify words as ‘‘aloud’’ or ‘‘silent’’. This fits well with the idea that the distinctive information that a piece of information was spoken aloud during study can be accessed at test to verify that prior study. It is this additional information, not greater strength, that underlies the production effect.

Manuscript received 11 December 2012
Manuscript accepted 24 April 2013
First published online 28 May 2013

REFERENCES

Batchelder, W. H., & Riefer, D. M. (1990). Multinomial processing models of source monitoring. *Psychological Review*, *97*, 548–564.

Batchelder, W. H., & Riefer, D. M. (1999). Theoretical and empirical review of multinomial process tree modelling. *Psychonomic Bulletin & Review*, *6*, 57–86.

Bayen, U. J., Murnane, K., & Erdfelder, E. (1996). Source discrimination, item detection, and multinomial models of source monitoring. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*, 197–215.

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

Bink, M. L., Marsh, R. L., & Hicks, J. L. (1999). An alternative conceptualisation to memory ‘‘strength’’ in reality monitoring. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*, 804–809.

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.

Castel, A. D., Rhodes, M. G., & Friedman, M. C. (2013). Predicting memory benefits in the production effect: The use and misuse of self-generated distinctive cues when making judgements of learning. *Memory & Cognition*, *41*, 28–35.

Cohen, R. L. (1981). On the generality of some memory laws. *Scandinavian Journal of Psychology*, *22*, 267–281.

Conway, M. A., & Gathercole, S. E. (1987). Modality and long-term memory. *Journal of Memory and Language*, *26*, 341–361.

Craik, F. I. M., & Lockhart, R. S. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning & Verbal Behavior*, *11*, 671–684.

Craik, F. I. M., & Tulving, E. (1975). Depth of processing and the retention of words in episodic memory. *Journal of Experimental Psychology: General*, *104*, 268–294.

Dodson, C. S., Prinzmetal, W., & Shimamura, A. P. (1998). Using Excel to estimate parameters from observed data: An example from source memory data. *Behavior Research Methods, Instruments & Computers*, *30*, 517–526.

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.

Donaldson, W. (1996). The role of decision processes in remembering and knowing. *Memory & Cognition*, *24*, 523–533.

Engelkamp, J. (1998). *Memory for actions.* Hove, UK: Psychology Press.

Engelkamp, J., & Krumnacker, H. (1980). Imaginale und motorische Prozesse beim Behalten verbalen Materials [Imagery and motor processes in memory of verbal material]. *Zeitschrift für experimentelle und angewandte Psychologie*, *27*, 511–533.

Fawcett, J. M. (2013). The production effect benefits performance in between-subject designs: A meta-analysis. *Acta Psychologia*, *142*, 1–5.

Forrin, N. D., Jonker, T. R., & MacLeod, C. M. (in press). Production improves memory equivalently following elaborative vs. non-elaborative processing. *Memory.*

Forrin, N. D., Ozubko, J. D., & MacLeod, C. M. (2012). Widening the boundaries of the production effect. *Memory & Cognition*, *40*, 1046–1055.

Franck, N., Phillippe, R., Dapriti, E., Dalery, J., Marie-Cardine, M., & Georgieff, N. (2000). Confusion between silent and overt reading in schizophrenia. *Schizophrenia Research*, *41*, 357–364.

Gathercole, S. E., & Conway, M. A. (1988). Exploring long-term modality effects: Vocalisation leads to best retention. *Memory & Cognition*, *16*, 110–119.

Hale, A. D., Hawkins, R. O., Sheeley, W., Reynolds, J. R., Jenkins, S., Schmitt, A. J., & Martin, D. A. (2011). An investigation of silent versus aloud reading comprehension of elementary students using Maze assessment procedures. *Psychology in the Schools*, *48*, 4–13.

Hintzman, D. L., Curran, T., & Oppy, B. (1992). Effects of similarity and repetition on memory: Registration without learning? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *18*, 667–680.

Hirst, W., Johnson, M. K., Kim, J. K., Phelps, E. A., Risse, G., & Volpe, B. T. (1986). Recognition and recall in amnesics. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *12*, 445–451.

Hirst, W., Johnson, M. K., Phelps, E. A., & Volpe, B. T. (1988). More on recognition and recall in amnesics. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*, 758–762.

Hockley, W. E. (2008). The effects of environmental context on recognition memory and claims of remembering. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*, 1412–1429.

Hopkins, R. H., & Edwards, R. E. (1972). Pronunciation effects in recognition memory. *Journal of Verbal Learning & Verbal Behavior*, *11*, 534–537.

Hourihan, K. L., & MacLeod, C. M. (2008). Directed forgetting meets the production effect: Distinctive processing is resistant to intentional forgetting. *Canadian Journal of Experimental Psychology*, *62*, 242–246.

Hunt, R. R. (2006). The concept of distinctiveness in memory research. In R. R. Hunt & J. Worthen (Eds.), *Distinctiveness and memory* (pp. 3–25). Oxford, UK: Oxford University Press.

Hunt, R. R. (2013). Precision in memory through distinctive processing. *Current Directions in Psychological Science*, *22*, 10–15.

Hunt, R. R., & Worthen, J. (Eds.) *Distinctiveness and memory.* Oxford, UK: Oxford University Press.

Johnson, M. K., Hashtroudi, S., & Lindsay, D. S. (1993). Source monitoring. *Psychological Bulletin*, *114*, 3–28.

Lin, O., & MacLeod, C. M. (2012). Aging and the production effect: A test of the distinctiveness account. *Canadian Journal of Experimental Psychology*, *66*, 212–216.

MacDonald, P. A., & MacLeod, C. M. (1998). The influence of attention at encoding on direct and indirect remembering. *Acta Psychologica*, *98*, 291–310.

MacLeod, C. M. (2011). I said, you said: The production effect gets personal. *Psychonomic Bulletin & Review*, *18*, 1197–1202.

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.

Ozubko, J. D., Gopie, N., & MacLeod, C. M. (2012). Production benefits both recollection and familiarity. *Memory & Cognition*, *40*, 326–338.

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.

Ozubko, J. D., & MacLeod, C. M. (2010). The production effect in memory: Evidence that distinctiveness underlies the benefit. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*, 1543–1547.

Paivio, A. (1971). *Imagery and verbal processes.* Oxford, UK: Holt, Rinehart & Winston.

Slamecka, N. J., & Graf, P. (1978). The generation effect: Delineation of a phenomenon. *Journal of Experimental Psychology: Human Learning and Memory*, *4*, 592–604.

Smith, S. M. (1979). Remembering in and out of context. *Journal of Experimental Psychology: Human Learning and Memory*, *4*, 460–471.

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

Yonelinas, A. P. (2002). The nature of recollection and familiarity: A review of 30 years of research. *Journal of Memory and Language*, *46*, 441–517.

Zimmer, H. D., Cohen, R. L., Guynn, M. J., Engelkamp, J., Kormi-Nouri, R., & Foley, M. A. (2001). *Memory for action: A distinct form of episodic memory?* New York, NY: Oxford University Press.