January 2020

This manuscript is a pre-publication version of an article published as:

Quinlan, C. K., & Taylor, T. L. (2019). Mechanisms underlying the production effect for singing. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale, 73*(4), 254–264

Abstract

The production effect is defined as better memory for items that were read aloud compared to items that were read silently. Quinlan and Taylor (2013) expanded the findings of the production effect by demonstrating that singing items produces even better recognition performance than reading aloud, and argued that this was due to enhanced relative distinctiveness. The current study tested three alternative accounts. In Experiment 1, we explored whether singing results in a larger production effect because it is deemed more bizarre than reading aloud. To address this, we tested a sample for whom singing does *not* seem bizarre: experienced singers. They also showed better recognition of items that were sung compared to those that were read aloud. In Experiment 2, we determined that singing appears to take longer than either reading aloud or reading silently; however, the possible effect of production time was further explored in Experiment 3. We did this by instructing participants to sing quickly, read aloud slowly, or read silently. Altering relative production times resulted in no discernable changes in subsequent recognition performance. Finally, in Experiment 4, we explored whether singing might strengthen the memory trace relative to reading aloud. We tested this by manipulating the production instruction between subjects. This eliminated the recognition advantage for both reading items aloud as well as for singing them aloud. Having ruled out these alternatives, we argue that singing improves subsequent recognition because it offers more distinctive elements than either reading aloud or reading silently.

Key words: Cognition; memory; production effect; distinctiveness; singing

Public significance statement: Compared to when words are read silently at study, subsequent recognition is better if those words are read aloud, and best if they are sung. We rule out other alternative explanations to suggest that singing is a particularly effective memory strategy because the words become relatively distinct.

Mechanisms Underlying the Production Effect for Singing

In 2010, MacLeod, Gopie, Hourihan, Neary, and Ozubko conducted an experiment that presented participants with a series of words, half of which participants were asked to read aloud and half of which they were asked to read silently. Following the study phase, participants were presented with an old/new recognition test. Recognition was significantly greater for words that were read aloud compared to words that were read silently; this difference in memory performance is known as the production effect (MacLeod et al., 2010; see MacLeod & Bodner, 2017 for review). This effect occurs for several types of production including reading aloud, whispering, mouthing, spelling, writing, typing (see Forrin, MacLeod, & Ozubko, 2012), and singing (Quinlan & Taylor, 2013), and influences recognition of both picture and word stimuli (Fawcett, Quinlan, & Taylor, 2012). While the effect consistently occurs in within-subject study designs, it is not commonly observed in between-subjects designs (MacLeod et al., 2010; although see Fawcett, 2013) and when it does occur in between-subjects designs, its magnitude is considerably smaller compared to within-subject designs.

Many researchers (e.g., Fawcett et al., 2012; Forrin et al., 2012; Forrin & MacLeod, 2016; Ozubko & MacLeod, 2010) have argued that the production effect is best explained by a distinctiveness account (although for alternative views, see Bodner & Taikh, 2012; Fawcett & Ozubko, 2016; and Jamieson, Mewhort, & Hockley, 2016). This account assumes that producing an item results in a relatively more "distinct" memory trace, making produced items easier to retrieve at test compared to non-produced items (see Schmidt, 1991, and Hunt, 2006, for reviews of distinctiveness). For example, MacLeod et al. (2010) and Ozubko and MacLeod (2010) have proposed that compared to reading silently, reading aloud consists of two additional distinct elements: articulation and audition, which are encoded at study. At test, participants can

use these two distinct elements as a retrieval cue: If participants remember saying the word aloud and/or hearing themselves say the word aloud, they can use that information heuristically to decide that the item was presented at study (see Schacter, Israel, & Racine, 1999). In a similar vein, other forms of production, such as spelling, typing, writing, and mouthing (see Forrin et al., 2012) are presumed to involve at least one additional element that is not present when reading silently and that can be used to advantage later retrieval.

Supporting a distinctiveness account, there is a systematic relation between the presumed number and strength of distinct elements and subsequent memory performance. For example, Fawcett and colleagues (2012) demonstrated an interaction between the production effect and the picture superiority effect, which they suggested arose because producing a word aloud resulted in at least two distinct elements (articulation and audition), whereas producing a picture aloud resulted in at least three distinct elements (articulation, audition, and visual detail). They argued that these distinct productions worked synergistically to produce a larger production effect for pictures than for words. Similarly, Forrin and colleagues (2012) showed that the magnitude of the production effect varied directly with the number of proposed distinct elements. Whereas they found a significant production effect for non-vocal forms of motor production, such as writing and mouthing, which contain one distinct element (motor movement), the magnitude of the effect was significantly larger for items read aloud, which contain two distinct elements (articulation and audition). Thus, in the literature, there seems to be a direct relation between the number of possible distinct elements inherent in processing an item and subsequent memory performance.

It is not the case, however, that reading aloud is a privileged form of production. This point was made clear in a study by Quinlan and Taylor (2013), who demonstrated that there are

other forms of verbal production, such as reading aloud loudly and singing, that have even more pronounced effects on subsequent memory compared to reading aloud in a normal voice (although see Hassall, Quinlan, Turk, Taylor, & Krigolson, 2016[1]). The fact that reading aloud loudly results in a larger production effect than reading aloud in a normal voice extends work that shows reading aloud in a normal voice results in a greater production effect than reading aloud in a whisper (Castel, 2009; Castel et al., 2013; Forrin et al., 2012). Together, these findings argue that distinct elements are defined not only qualitatively (i.e., the type of distinct elements) but also quantitatively (the intensity of those elements). To wit: Reading aloud loudly is more distinctive than reading aloud in a normal voice because of the increased volume of the production; singing aloud is more distinctive than reading aloud because singing includes unique additional elements not available during reading (e.g., pitch, melody, timbre).

Thus, the observation that singing results in a greater production effect than either reading aloud loudly or reading aloud in a normal voice is consistent with a distinctiveness account of the production effect. There are, however, alternative explanations that have not yet been ruled out. Testing these explanations is the goal of this study. In Experiment 1, we ruled out a bizarreness account. According to the bizarreness hypothesis, the greater production effect for singing versus reading aloud could be due to the simple fact that singing is an unusual or bizarre action for many people. Our results countered this hypothesis by demonstrating that experienced singers – for whom singing is *not* a bizarre or unusual activity – replicated Quinlan and Taylor's

[1] In contrast to Quinlan and Taylor (2013), Hassall et al. (2016) found that the production effect for singing was not significantly greater than the production effect for reading aloud. The experiment conducted by Hassall et al. (2016) incorporated electroencephalography (EEG) at study and as a result, there were substantial methodological differences between the experiment conducted by Hassall et al. (2016) and the three experiments reported in Quinlan and Taylor (2013), which can likely account for the different findings. These differences are discussed in further detail by Hassall et al. (2016).

(2013) larger production effect for items sung compared to items read aloud. In Experiment 2, we determined that singing might take more time to produce than reading aloud and explored whether this additional production time might be responsible for better subsequent recognition, along the lines of the total time hypothesis (Cooper & Pantle, 1967). In Experiment 3, we put this hypothesis to the test by instructing participants to sing *quickly* and to read aloud *slowly*. Even though this explicit instruction might not have been sufficient to completely eliminate the longer production durations for items sung compared to those read aloud, it should have nevertheless *reduced* those differences. Yet the production effect continued to be larger for items sung than for items read aloud, with no discernible difference compared to the results of Quinlan and Taylor (2013) who did not explicitly instruct participants to control production durations. Finally, in Experiment 4, we tested whether the memory representations of items sung might be stronger than the memory representations of items read aloud. We hypothesized that if singing produces a stronger memory trace than reading aloud, we would find a significantly greater production effect for items sung compared to those read aloud even in a between-subjects design (i.e., because a strong encoding would prevail whether a subset of items was sung or all items were sung). However, in our between-subjects manipulation, neither singing nor reading aloud yielded a significant production effect, which counters a pure strength-based account.

Experiment 1

Quinlan and Taylor (2013) found that the production effect was greater when participants sang items compared to when they read items aloud. Although it was argued that this difference in the magnitude of the production effect was a result of distinctiveness, another possibility is that singing may be due to a bizarreness effect. A bizarreness effect refers to better subsequent memory performance for information (i.e., words or sentences) or for self-performed or

imagined actions (e.g., Engelkamp, Zimmer, & Biegelmann, 1993; Engelkamp, Zimmer, Mohr, & Sellen, 1994; Mohr, Engelkamp, & Zimmer, 1989; Worthen & Wood, 2001) that are perceived as bizarre or unusual rather than common (McDaniel, Anderson, Einstein, & O'Halloran, 1989; McDaniel & Einstein, 1986; McDaniel, Einstein, DeLosh, May, & Brady, 1995; for review, also see McDaniel & Bugg, 2008). There have been several theoretical explanations for the bizarreness effect (e.g., expectation violation framework, incongruity theory) that share the common assumption that items deemed bizarre tend to violate the perceiver's expectations and hence to generate surprise; this reaction focuses attention and cognitive control mechanisms, ultimately enhancing both contextual encoding and item elaboration (e.g., Hirshman, Whelley, & Palij, 1989; Mather & Carstensen, 2005; Schmidt, 1991).

Although bizarreness and distinctiveness both result in better subsequent memory, they do so via different mechanisms. In a distinctiveness framework, the distinct elements of items that are encoded at study are later used as retrieval cues at test. In contrast, the bizarreness account suggests that it is the strength of items at encoding and any associated contextual cues/information that functions to enhance memory performance at test. With regard to enhanced memory performance, the distinctiveness account focuses on the interaction of processes at encoding and retrieval, whereas the theoretical accounts associated with a bizarreness effect focus only on the processes that occur at encoding. Furthermore, theoretical accounts associated with the bizarreness effect emphasize that bizarreness is defined relative to the individual, whereas distinctiveness is defined as relative to the context. Based on a distinctiveness account, in a mixed-list, within-subject study design, items that are sung will be remembered

better because they are in the context of other less distinct items (i.e., items read aloud, items read silently). In contrast, based on a bizarreness account, in a mixed-list, within-subject study design, sung items will be remembered better only to the extent that singing is perceived to be bizarre by the individual participant.

The goal of Experiment 1 was to determine whether the greater production effect for singing versus reading aloud could be attributed to the fact that singing in Taylor and Quinlan's (2013) task might have been perceived by their participants as unusual or "bizarre". To address this, we replicated the methodology of Quinlan and Taylor (2013) but required that all participants have at least one year of singing experience. Although it is conceivable that singing individual words (as opposed to an entire phrase or song) may be bizarre regardless of singing experience, our rationale was that the act of singing individual words should be substantially *less* bizarre for participants who have a history of singing experience. Thus, if a greater production effect for singing versus reading aloud is attributable to a bizarreness effect, this difference should be eliminated or reduced in a sample of experienced singers.

In a within-subject design, experienced singers were presented with a study phase followed by a test phase. In the study phase, words appeared one at a time in one of three coloured fonts with each colour representing a particular instruction condition: Sing, Read Aloud, and Read Silently. Immediately following the presentation of all study items, these experienced singers completed a yes/no recognition test that included all study items as well as an equal number of coloured foil items that had not been presented at study.

Method

Participants. Sixteen students participated in this experiment in exchange for credit toward their grade in an eligible Psychology class at Dalhousie University or for $10.00 compensation. Each

participant was tested in one session lasting approximately 30 minutes. All participants reported normal or corrected-to-normal vision and a good understanding of the English language. Furthermore, all participants were required to have a minimum of one year of singing experience as a performer, member of a choir, or equivalent. Participants reported between 2 and 20 years of singing experience, with a mean of 8.19 years (*SD*= 6.27), most often in choirs. All 16 participants reported that singing was a normal, comfortable activity for them that they engaged in daily.

Stimuli and Apparatus. This experiment was run using PsyScope X (cf. Cohen, MacWhinney, Flatt, & Provost, 1993) loaded on a 24" iMac computer running Mac OSX Leopard, version 10.5. All responses were recorded on a standard Macintosh Universal Serial Bus keyboard. All text was presented against a uniform black background in Time New Roman size 42 font.

We used the same 240-word list employed by Taylor and Quinlan (2013). Prior to running each participant, custom software randomly distributed the 240 words into six lists of 40 words each. This ensured a unique list composition for each participant. Three of these lists were assigned to the Sing, Read Aloud, and Read Silently conditions of the study phase. According to the mapping of colour to production instruction, the words on one of these lists were always presented in red font; the words on another were always presented in blue font; and the words on the other were always presented in white font. The remaining three lists were designated as unstudied foil items to be presented during the recognition test only. The words on one of these lists were always presented in red; the words on another list were always presented in blue; and the words of the other list were always presented in white.

Procedure. Prior to beginning the experiment, information related to participant's singing experience was collected. Each participant was asked the following three questions: 1) How

many years of singing experience do you have? 2) What type of singing experience do you have? 3) Would you consider singing to be a comfortable, everyday task for you?

Following this oral questionnaire, participants were given oral instructions, which were later re-iterated on the computer monitor. The experimenter told participants that in the study phase they would see a series of words printed in red, blue, or white. Approximately one third of participants (n=6) were told to sing the words in red, read aloud the words in blue, and read silently the words in white; approximately one third of participants (n=5) were told to sing the words in blue, read aloud the words in white, and read silently the words in red; and approximately one third of participants (n=5) were told to sing the words in white, read aloud the words in red, and read silently the words in blue. The experimenter told participants that instructions for a memory test would appear on screen following the presentation of all study words. The nature of the memory test (i.e., recognition, recall) was not specified to participants.

Before beginning the experiment proper, the experimenter told participants that they would be presented with a familiarization phase and a practice phase that were designed to ensure that they were comfortable with the three instruction conditions (Sing, Read Aloud, Read Silently). For the Sing condition, participants were instructed to sing as they would in any other context (e.g., in the car, in the shower) and thus the singing strategies used by participants varied depending on their individual style. The experimenter remained in the room with participants until the end of the practice phase, to ensure that participants followed the instructions for the three conditions (Sing, Read Aloud, Read Silently). Based on participant observation during the familiarization phase, participants applied similar pitch, melody, and timbre to the words that they were instructed to sing; however, this was not monitored or measured during the study phase. We presume that the uniformity of their production technique allowed participants to

concentrate their efforts on remembering the words, rather than on trying to generate a unique combination of pitch, melody, and timbre for each item that needed to be sung. In that respect, there was no sense in which participants exhibited more trial-to-trial differences in their sing productions than they did in their read aloud productions.

Familiarization phase. Prior to beginning the study phase, participants were presented with 15 trials. These trials were designed to familiarize participants with the three font colors (red, blue, white) and their associated instruction (Sing, Read Aloud, Read Silently). Five trials of each color were intermixed randomly. On each familiarization trial, a blank screen was presented for 500 ms followed by the verbal descriptor of the color along with its associated production instruction (e.g., 'RED-Sing'). Both the name of the color as well as the associated production instruction were printed in the indicated color (e.g., 'RED-Sing' was printed in red), centred on the computer monitor, and displayed for 2000 ms.

Practice phase. Immediately following the familiarization phase, participants completed a practice phase. On each practice trial, a blank screen was presented for 500 ms followed by the word 'banana' at centre for 2000 ms. The word 'banana' was printed in one of three colors (red, blue, white), which corresponded to one of three conditions (Sing, Read Aloud, Read Silently). Five trials of each condition were intermixed randomly to produce a total of 15 practice trials. These practice trials were identical to those in the study phase with the exception that 'banana' was the only word presented. This phase was designed to give participants practice with the three production conditions.

Study phase. Immediately following the last trial in the practice phase, the study phase trials began. The study phase consisted of a total of 120 trials. There were 40 trials in each of the three conditions (Sing, Read Aloud, Read Silently), which were intermixed randomly. Each

study trial began with a blank screen for 500 ms followed by a word in the centre of the computer monitor for 2000 ms. Each word was selected randomly without replacement from the Sing, Read Aloud, and Read Silently study lists and coloured accordingly. The total duration of each study trial was 2500 ms.

Recognition phase. Upon completing the study phase, participants began the recognition phase. The recognition phase consisted of a self-paced yes/no recognition test. At the beginning of the recognition phase, instructions were presented at the top of the computer screen. Participants were instructed to press the 'y' key if they recognized the word from the study trials and to press the 'n' key if they did not recognize the word from the study trials (i.e., a foil word). All responses could be self-corrected using the backspace key and were submitted by pressing the space bar. The next recognition trial began after a response was submitted.

To calculate separate foil false alarm rates for each condition, we maintained the colour coding from the study phase such that items that were studied in red were tested in red; items that were studied in blue were tested in blue; and items that were studied in white were tested in white. As already noted, equal numbers of foil items were presented likewise presented in red, blue, and white (the same approach also used by Quinlan & Taylor, 2012; see also Bodner, Taikh, & Fawcett, 2013). Thus, on each recognition trial, one word printed in red, blue, or white colored font was presented at the centre of the computer monitor until a response was made. In total, there were 240 recognition trials comprised of the randomly intermixed presentation of the 40 Sing, 40 Read Aloud, and 40 Read Silently study words, and the 40 red, 40 blue, and 40 white unstudied foil words. Although the items in the yes/no recognition phase were printed in different colors, participants were not instructed to perform the condition (Sing, Read Aloud, Read Silently) corresponding to the color of the item.

Data Analysis. All analyses were conducted using R (R Core Team, 2018; Singmann, 2018) via jamovi software (The jamovi project, 2019).

Results

A hit was defined as a 'yes' response to studied words from the Sing, Read Aloud, and Read Silently conditions; a false alarm was defined as a 'yes' response to unstudied foil words. The false alarms were classified according to the meaning of the colour coding at study (e.g., if a participant responded "yes" to the recognition of a foil word printed in red and red had signaled the Sing condition at study, the foil response was considered to be a false alarm for the Sing condition). The mean proportions of hits and foil false alarms are shown in Table 1.

Insert Table 1 about here

On a subject-by-subject basis, a corrected hit rate was calculated separately for each production condition by subtracting the proportion false alarms from the corresponding hit rate for that condition. The mean corrected hit rate was used to assess recognition memory performance and was analyzed in a one-way analysis of variance (ANOVA), with production condition (Sing, Read Aloud, Read Silently) as a within-subject factor. This analysis revealed an overall significant effect of production condition, $F(2, 30)=23.2$, $MSe=.008$, $p<.001$, $\eta^2=.362$. A clear pattern of results emerged: More items in the Sing condition were recognized than items in the Read Aloud condition, $t(15)=3.36$, $p=.004$, and more items in the Read Aloud condition were recognized than items in the Read Silently condition, $t(15)=3.83$, $p=.002$.

Discussion

The goal of Experiment 1 was to determine whether the greater production effect for singing versus reading aloud could be explained as a bizarreness effect (Hirshman et al., 1989;

Mather & Carstensen, 2005; Schmidt, 1991). To test this, we replicated the methodology of Quinlan and Taylor (2013) but required that participants have a history of singing experience so that singing would not be deemed unusual. We predicted that if the Sing > Read Aloud > Read Silently pattern of memory performance observed in Quinlan and Taylor (2013) was due to singing being perceived as a "bizarre" or unusual act (Hirshman et al., 1989; Mather & Carstensen, 2005; Schmidt, 1991), then relative to the memory advantage for reading aloud over reading silently, the memory advantage for singing over reading aloud would be reduced or eliminated for participants with singing experience. In contrast to this prediction, the current experiment replicated our previous findings, showing a significant production effect for the Sing and Read Aloud conditions, with greater memory performance in the Sing condition than in the Read Aloud condition. Indeed, the results were indistinguishable from those of Quinlan and Taylor (2013; Experiment 2). To enable this comparison, we used corrected hit rates to calculate a production effect by subtracting recognition performance in the Read Silently condition from that of the Read Aloud condition and the Sing condition. For the Sing condition, the magnitude of this production effect was 0.220 in the current study and 0.229 in Quinlan and Taylor's (2013) study [$t(34)=.184$, $p=.855$]; for the Read Aloud condition, the magnitude of this effect was .129 in the current study and .109 in Quinlan and Taylor's (2013) study [$t(34)=.426$, $p=.673$]. Although these results do not provide direct evidence in support of the distinctiveness account, they are inconsistent with a bizarreness account.

Experiment 2

Another alternative explanation for the greater production effect for singing versus reading aloud relates to production duration. It is conceivable that singing items usually takes longer than reading items aloud or silently. To the extent that a longer production duration

translates into greater processing (see Cooper & Pantle, 1967, for review), this could account for better recognition of items sung compared to items read aloud. To test this, we required participants in Experiment 2 to press the computer mouse key when they began to sing, read aloud, or read silently, and to release the mouse key when they were finished. This provided an estimate of the relative duration of these three productions and allowed us to determine whether singing does, in fact, take longer to produce than reading aloud. Importantly, using a subjective report of production duration allowed us to estimate the baseline duration for reading silently, which required no overt response and was therefore not amenable to an objective measurement, such as a voice key. This allowed us to determine whether, in turn, reading aloud takes longer than reading silently.

Even allowing for a potentially imperfect relation between reported and true production duration, we had no *a priori* reason to presume that temporal resolution would differ for the three forms of production. Obtaining a self-reported estimate of production duration would therefore allow us to determine the relative ranking of the time needed to sing, read aloud, and read silently. To the extent that production time is primarily responsible for subsequent memory performance, the relative ranking of production time should mimic that of recognition performance, such that: Singing > Read Aloud > Read Silently.

Method

Participants. Thirty-six undergraduate students participated in this experiment in exchange for credit toward their grade in an eligible Psychology class at Dalhousie University. This sample size was larger than Experiment 1 due to the expectation that measuring production duration would require additional power. Each participant was tested in one session lasting approximately

30 minutes. All participants reported normal or corrected-to-normal vision and a good understanding of the English language.

Stimuli and Apparatus. To accommodate the desire to detect both the press and the release of the computer mouse, we elected to use SuperLab, loaded on a 27" iMac computer. The same 240-item word list used for Experiment 1 was divided randomly into 6 lists of 40 items each. The monitor background remained a uniform white. For this reason, the colours used as instructions were changed relative to Experiment 1. Two lists were assigned a Red coloured font, two were assigned a Blue coloured font, and two were assigned a Black coloured font. The assignment of colour to production instruction was counterbalanced across participants. For half of the participants within each colour mapping, the first list in each pair served as the study list and the second list as unstudied foils; this list designation was reversed for the other half of the participants. Our sample size was a multiple of 12 to ensure equal numbers of participants in each of the counterbalancing conditions. All study and test words were presented at the centre of the computer monitor, in size-20 Geneva font.

Procedure. Study and test words were presented in Red, Blue, or Black, with production condition mapped to colour according to the counterbalancing condition. The instructions emphasized not only the production that was required but also the need to depress the mouse button at the start of the production and to release the mouse button at the end of the production.

Practice phase.

After receiving detailed instructions, participants were given practice trials. Each trial started with black fixation crosshairs ("+") presented in the centre of the computer monitor for 500 ms, followed by the word "word" printed in Red, Blue, or Black. The participant was required to produce this word according to the production instruction while also depressing the

mouse button to report the start of the production and releasing the mouse button to report the end of the production. Each "word" was presented for 2000 ms followed by a 2000 ms blank screen; the mouse was monitored for this full 4000 ms duration. If participants failed to depress or release the mouse button during these practice trials, they received an onscreen alert that reminded them of the importance of using the mouse to report the start and end of each production – even when the production was reading silently.

Study phase. The study phase was as described for Experiment 1, with the following exceptions. Each study phase trial began with fixation crosshairs ("+") printed in black in the middle of the computer monitor for 500 ms. This was replaced by a study word for 2000 ms; this word was coloured Red, Blue, or Black, according to the list from which it was drawn. During the word presentation and for 2000 ms following, the mouse was monitored for press/release actions.

Recognition phase. The recognition phase was as described for Experiment 1, except that a press of the return key (rather than the space bar) was used to enter each response and advance to the next trial.

Results

The data contributed by one participant were removed from all analyses, due to reported production durations that exceeded by almost 3 standard deviations the mean of all other participants. The mean proportions of hits and false alarms for the 35 remaining participants were calculated in the same manner as described in Experiment 1 and are shown in Table 1.

The mean corrected hit rate was analyzed in a one-way analysis of variance (ANOVA) with production condition (Sing, Read Aloud, Read Silently) as a within-subject factor. This analysis revealed an overall significant effect of production condition, $F(2, 68)=20.9$, $MSe=.012$,

$p<.001$, $\eta^2=.130$. Similar to Experiment 1, more items in the Sing condition were recognized than items in the Read Aloud condition, $t(34)=2.67$, $p=.012$, and more items that were Read Aloud were recognized than items that were Read Silently, $t(34)=3.94$, $p<.001$.

The average reported production duration was likewise analyzed in a one-way ANOVA with production condition (Sing, Read Aloud, Read Silently) as a within-subject factor. This analysis revealed a significant effect of reported production condition, $F(2, 68)=42.9$, $MSe=32245$, $p<.001$, $\eta^2=.142$. The mean reported production duration was significantly longer in the Sing condition ($M=1194$ ms, $SD=399$) than in the Read Aloud condition ($M=878$ ms, $SD=354$; $t(34)=7.82$, $p<.001$), and the Read Silently condition ($M=826$ ms, $SD=456$; $t(34)=7.15$, $p<.001$). Interestingly, however, the mean reported production duration did not differ significantly in the Read Aloud and Read Silently conditions, $t(34)=1.46$, $p=.154$).

Discussion

The recognition results of Experiment 2 replicate the pattern reported by Quinlan and Taylor (2013), as well as the pattern of results in Experiment 1 of the current manuscript: Sing > Read Aloud > Read Silently. Nevertheless, the reported production durations showed a different pattern, such that: Sing > (Read Aloud = Read Silently). Parsimony favours an interpretation that the production duration follows a different pattern than the recognition data because the former does not determine the latter. Indeed, if it were the case that production duration was an important determinant of memory performance for items sung but not for items read aloud (or, for that matter, those read silently), this should be obvious by regressing recognition performance on production durations. However, when we calculated bivariate correlations to relate reported production durations to the corrected hit rates, there was no hint of a relation in *any* of the three production conditions: Sing (r=.134, $p=.442$), Read Aloud (r=.086, $p=.622$) or

Read Silently (r=.192, *p*=.269) – even allowing for the relatively small sample size. This suggests that production duration is not likely the primary determinant of the production effect in recognition.

Even though the results of Experiment 2 show no direct evidence of production duration as a primary determinant of the production effect, there were nevertheless significantly longer self-reported production durations for items that were sung versus items that were read aloud. As such, there is a possibility that a longer production duration *might* contribute indirectly to the relatively larger production effect for items sung than for items read aloud – even if it plays no apparent role in the production effect for items read aloud (versus silently). This led us to conduct another experiment to explore the potential role of production duration in the difference that occurs in recognition performance for items sung versus those read aloud.

Experiment 3

The goal of Experiment 3 was to further test whether the greater production effect for singing versus reading aloud can be attributed to differences in production duration. To address this, we made use of participants' apparent subjective awareness of their own production durations to instruct them to sing *quickly* and to read aloud *slowly*. We did not, however, independently verify production durations. This is because it was not immediately obvious what the best method would be for doing so – especially to the extent that we again wished to compare singing to both reading aloud and reading silently. On the one hand, objective measurements cannot be readily obtained for both overt (sing, read aloud) and covert (read silently) forms of production, given that the latter provide no measurable response. On the other hand, use of subjective reports as described in Experiment 2 allows for the measurement of both overt (singing, reading aloud) and covert (reading silently) productions. But we were concerned that

responses using the mouse could be readily contaminated by demand characteristics when control over production duration was an explicit requirement of the study: Knowing that they were supposed to sing *quickly* and to read aloud *slowly* might have induced participants to under-estimate their singing productions and over-estimate their reading productions.

In any case, we did not believe that measuring production times was critical to our purpose. We assumed that participants would have sufficient control over and insight into their productions to heed the instructions. Compared to Quinlan and Taylor (2013), who made no such demands on the production duration, it seemed reasonable that these instructions would encourage participants to sing more quickly than they might otherwise and to read more slowly than they might otherwise – even if singing did not become an overall faster production than reading aloud. Implementation of these instructions to sing *quickly* and read aloud *slowly* should be sufficient to reduce the memory advantage otherwise observed for singing – but only if production time is a critical determinant of subsequent recognition performance.

Method

Participants. A total of 43 students participated in this experiment in exchange for credit toward their grade in an eligible Psychology class at Dalhousie University. Despite the similar design, we intentionally used a larger sample size than in Experiment 1 to accommodate the change in production instructions, which we thought might be more challenging to implement due to the temporal qualifier (i.e., to sing *quickly* and to read *slowly*). Our goal was to collect data from 40 participants, but advance scheduling resulted in a total of 43; no data were analyzed prior to collecting data from the final scheduled participants. Each participant was tested in one session lasting approximately 30 minutes. All participants reported normal or corrected-to-normal vision and a good understanding of the English language.

Stimuli and Apparatus. The stimuli and apparatus were identical to Experiment 1.

Procedure. The general procedure was identical to Experiment 1, with the exception that the production instructions were to sing *quickly*, read aloud *slowly*, and read silently. For the Sing Quickly condition, participants were instructed to sing at a faster-than-normal pace and for the Read Slowly condition, participants were instructed to read aloud at a slower-than-normal pace. It was assumed that individual participants had insight into their "normal pace" sing and read aloud baselines and were able to exhibit self-control to adjust their pace according to instructions (i.e., quickly, slowly). The experimenter remained in the room during the familiarization and practice phases to ensure that participants understood and followed the instructions.

Results

The mean proportions of hits and false alarms were calculated in the same manner as described in Experiment 1 and are shown in Table 1.

The mean corrected hit rates were analyzed in a one-way analysis of variance (ANOVA) with production condition (Sing Quickly, Read Aloud Slowly, Read Silently) as a within-subject factor. This analysis revealed an overall significant effect of production condition, $F(2, 84)=64.5$, $MSe=.013$, $p<.001$, $\eta^2=.352$. Consistent with the previous two experiments, more items in the Sing Quickly condition were recognized than items in the Read Aloud Slowly condition, $t(42)=2.82$, $p=.007$, and more items in the Read Aloud Slowly condition were recognized than items in the Read Silently condition, $t(42)=7.35$, $p<.001$.

Discussion

The goal of Experiment 3 was to determine whether the greater production effect for singing versus reading aloud could potentially be explained by differences in relative production duration. Our rationale was that instructing participants to sing *quickly* and to read aloud *slowly*

would reduce the production duration difference that otherwise occurs in the absence of a production time instruction (see Experiment 2). The question was whether this reduction in production duration would result in a concomitant reduction in the difference in recognition performance for items sung compared to those read aloud.

Consistent with our previous findings (Quinlan & Taylor, 2013) and those of Experiments 1 and 2, we found overall better subsequent recognition for items that were sung at study rather than read aloud. To determine whether the magnitude of this difference was affected by the instructions to control the production rates for items sung and for items read aloud, compared memory performance in the current experiment (Sing Quickly, Read Aloud Slowly, Read Silently) to that of Quinlan and Taylor's (2013) Experiment 2 (Sing, Read Aloud, Read Silently). We analyzed corrected hits in an ANOVA, with production condition (Sing/Sing Quickly, Read Aloud/Read Aloud Slowly, Read Silently) as a within-subject factor and experiment as a between-subjects factor[2]. Neither the main effect of experiment ($F<1$) nor the two-way interaction between production condition and experiment was significant ($F(2,122)=2.55$, $MSe=0.012$, $p=.082$). Thus, despite our instruction to participants to control their speed of production, there was no significant difference between the pattern of results in the current experiment and those obtained by Quinlan and Taylor (2013). This suggests that a difference in production duration is unlikely to account for the greater production effect that occurs for items that are sung compared to those that are read aloud.

[2] The results of the current experiment were compared to those of Quinlan and Taylor (2013; Experiment 2), rather than those of Experiment 2 of the current manuscript due to consistency in methodology between the current experiment and Quinlan and Taylor (2013; Experiment 2). Experiment 2 of the current manuscript involved a longer study trial duration (i.e., 4500ms as opposed to 2500ms) and an additional manipulation (i.e., self-reported production durations), and thus, was not the most suitable experiment for comparison.

That said, a difficulty with our interpretation arises from the fact that we did not independently verify production times. Had the instruction to control production duration produced an effect on subsequent recognition (even if not fully accounting for *differences* in recognition), we would have a stronger claim for the effectiveness of the instruction. As it is, the result is subject to the criticism that participants might not have followed the instructions to control their rates of production. This concern is underscored by the finding that there was no significant difference between the pattern of results in the current experiment and that of Experiment 2 in Quinlan and Taylor (2013): Reading aloud at a presumed slower-than-normal pace produced no discernible improvement in subsequent memory performance and singing at a faster-than-normal pace produced no discernible reduction in subsequent memory performance.

Despite this, we are inclined to argue against an important role for production duration. The results of Experiment 2 did not provide convincing evidence of a relation between production duration and subsequent recognition for any of the production conditions. And instructing participants to change the speed of production in Experiment 3 did not produce any discernable changes in the pattern of subsequent recognition performance for items that were sung versus those that were read aloud. This does not mean that production duration plays no role in the greater production effect for items that are sung versus those that are read aloud. But in the absence of compelling evidence to the contrary, parsimony favours the interpretation that singing aloud and reading aloud affect subsequent memory due to the operation of a common underlying mechanism. There is no good reason to believe that distinctiveness alone affects recognition for items that are read aloud but that distinctiveness and duration combine to improve recognition for items that are sung.

Experiment 4

A distinctiveness account of the production effect claims that produced items are only distinct in relation to a backdrop of non-produced items; when there is no backdrop of non-produced items, produced items are no longer distinct (MacLeod et al., 2010). In within-subject designs, participants are presented with both produced and non-produced items during study, which tend to be intermixed randomly. This type of design allows for produced items to be processed distinctively in relation to a contextual backdrop of non-produced items. In contrast, in between-subjects designs, participants are presented with either only produced or only non-produced items. Because this type of study design does not provide any contextual or relational information between produced and non-produced items during study, neither of the item sets should be processed distinctively (although see Jonker, Levene, & MacLeod, 2013). Thus, a distinctiveness account predicts a significant production effect in within-subject designs, but not in between-subjects designs.

An alternative to a distinctiveness account of the production effect is a pure strength-based account. A pure strength-based account assumes that, compared to non-produced items, produced items are processed and encoded more elaborately at study, thereby resulting in a stronger memory representation (e.g., Bodner & Taikh, 2012). In contrast to a distinctiveness account, a pure strength-based account predicts that produced items should be processed and encoded elaborately at study regardless of whether they are presented alone or in the context of non-produced items. In other words, a pure strength-based account predicts a production effect regardless of whether the production conditions are presented in a within-subject design or a between-subjects design.

Testing this prediction, a pure strength-based account of the production effect has largely been ruled out with respect to the improved memory performance that occurs for items read aloud versus silently (e.g., Dodson & Schacter, 2001; Hopkins & Edwards, 1972; MacLeod et al., 2010; but see Bodner & Taikh, 2012; Fawcett & Ozubko, 2016; and Jamieson et al., 2016). Parsimony again favours the prediction that the production effect for singing aloud should likewise counter a strength-based account, under the assumption that the production effect for singing should operate according to the same underlying principles as the production effect for reading aloud. However, this remains an empirical question. It is possible that greater recognition of items that are sung compared to those that are read aloud might be due to a combination of distinctiveness and increased memory trace strength.

If singing items results in stronger memory traces than reading items aloud and reading items silently, this should be evident in a between-subjects design. The goal of Experiment 4 was to test this possibility.

Method

Participants. Sixty undergraduate students participated in this experiment in exchange for credit toward their grade in an eligible Psychology class at Dalhousie University. Twenty participants were assigned to each of the three between-subjects conditions: Sing, Read Aloud, and Read Silently. Each participant was tested in one session lasting approximately 30 minutes. All participants reported normal or corrected-to-normal vision and a good understanding of the English language.

Stimuli and Apparatus. As in MacLeod et al. (2010; Experiment 2), all study words were presented in white colored font and all yes/no recognition words were printed in yellow colored font. Participants were only exposed to one type of production condition (e.g., Sing or Read

Aloud or Read Silently). It was therefore not necessary to use coloured fonts (i.e., red, blue, white) to differentiate the production conditions (or foils). Otherwise, the stimuli and apparatus were identical to Experiment 1.

Procedure. Participants were assigned to one of three instruction conditions — Sing, Read Aloud, or Read Silently. The experiment instructions varied accordingly across conditions. Participants in the Sing instruction condition were told that they should sing all study words aloud; participants in the Read Aloud instruction condition were told that they should read all study words aloud; and, participants in the Read Silently condition were told that they should read all study words silently (with no mouth movement or overt vocalization). The experimenter told participants that they would be required to complete a memory test following the presentation of all study words. There were no familiarization or practice phases. Other than the uniform white colour coding of study words and the uniform yellow colour coding of test words, the study and test phases were otherwise identical to Experiment 1.

Results

The mean proportions of hits and false alarms were calculated as described for Experiment 1 and are shown in Table 1.

The mean corrected hits were analyzed in a one-way analysis of variance (ANOVA) with production condition (Sing, Read Aloud, Read Silently) as a between-subjects factor. This analysis revealed no significant difference in proportion hits as a function of production condition, $F<1$. Indeed, there was no significant difference in recognition performance in the Sing condition compared to the Read Aloud condition ($t(38)=0.830$, $p=.411$) or in the Read Aloud condition compared to the Read Silently condition ($t(38)=1.34$, $p=.187$).

Discussion.

Experiment 4 used a between-subjects study design to determine whether differences in memory strength could account for the greater production effect for singing versus reading aloud. Based on previous findings using a between-subjects design (e.g., Dodson & Schacter, 2001; Hopkins & Edwards, 1972; MacLeod et al., 2010), we predicted that if singing induces a greater production effect than reading aloud due to differences in trace memory strength, we would find a significant between-subjects production effect for items that are sung compared to items that are read aloud or read silently. We did not. Our findings for the Read Aloud condition replicate the existing literature to demonstrate that the production effect (for reading aloud versus silently) tends to be limited by a between-subjects design (e.g., Dodson & Schacter, 2001; Hopkins & Edwards, 1972; MacLeod et al., 2010; although see Fawcett, 2013) consistent with a distinctiveness account (see MacLeod et al., 2010). The results of the present experiment also demonstrate that the larger production effect for singing versus reading aloud is likewise limited by a between-subjects design. This strongly suggests that singing results in better subsequent recognition than reading aloud for the same reason that reading aloud results in better subsequent recognition than reading silently: distinctiveness.

General Discussion

The literature has overall supported the interpretation that the production effect for reading aloud versus silently is due to relative distinctiveness of the production. It was incumbent upon us to determine whether the especially large production effect that Quinlan and Taylor (2013) observed for singing was likewise attributable to distinctiveness alone. In the current study, we questioned whether bizarreness, production duration, and memory strength contributed to the greater production effect that occurs for singing aloud versus reading aloud.

Using experienced singers, Experiment 1 replicated the pattern of recognition results reported by Quinlan and Taylor (2013): Sing > Read Aloud > Read Silently. Under the assumption that experienced singers would not find singing to be particularly unusual or bizarre, this finding counters the suggestion that Quinlan and Taylor (2013) observed a larger production effect for singing than for reading aloud only because participants deemed singing to be "bizarre". In contradicting a bizarreness account, our findings accord with those of Forrin et al. (2012) who observed a *smaller* production effect for writing, whispering, and mouthing, compared to reading aloud – despite the fact that whispering and mouthing are arguably more unusual/less frequent activities than either writing or reading aloud. Together, our Experiment 1 findings and those of Forrin et al. (2012) provide no support for a bizarreness explanation of the production effect.

Experiment 2 suggested that singing requires more time than reading aloud but that reading aloud and reading silently are not discernably different. This raised the possibility that the longer production time needed for singing might be responsible for the larger production effect compared to reading aloud, even if it plays no apparent role in the production effect for reading aloud. There was, however, no hint of a relation between production duration and recognition memory performance for *any* of the production conditions. Moreover, when we asked participants in Experiment 3 to explicitly control the duration of their productions – with the goal of reducing the difference in production duration for items sung versus items read aloud – there was no discernable effect: Subsequent recognition memory was better for items sung quickly than for items read aloud slowly, and this difference was indistinguishable from that

reported by Quinlan and Taylor (2013) who made no such demands for participants to control production duration.

In arguing against a critical role for production duration, our results again accord with those of Forrin et al. (2012). Forrin and colleagues (2012) reported a pattern of memory performance such that: Read Aloud > Write > Read Silently, despite the fact that writing likely takes longer than reading (either aloud or silently). In a similar vein, reading aloud loudly, reading aloud in a normal voice, and whispering likely all take approximately the same amount of time. Nevertheless, the combined recognition results of Quinlan and Taylor (2013) and those of Forrin et al. (2012) show that: Reading Aloud Loudly > Read Aloud > Whispering > Read Silently. Thus, even without objective measurement of the production time, there is no convincing evidence that production time plays an important role in the production effect for singing or for other vocal or non-vocal forms of production.

In Experiment 4, we used a between-subjects design to test whether singing might increase item strength, over and above any effects that might be attributable to relative distinctiveness. In this between-subjects design, there was no discernible evidence of a production effect either for items read aloud or for items sung. This counters the suggestion that singing aloud, in particular, might strengthen the item representation. In so doing, the results of Experiment 4 likewise bolster the conclusions of Experiments 1-3. To wit: If it were the case that singing seemed "bizarre" and thereby attracted additional attentional resources, the increased processing would be expected to increase the item strength; likewise, if the longer production duration for

singing items compared to reading items aloud enabled greater processing, this would also be manifest as increased trace strength.

The fact that our between-subjects manipulation provided no evidence of a production effect – for either reading aloud or for singing – is consistent with other published findings (e.g., Dodson & Schacter, 2001; Hopkins & Edwards, 1972; MacLeod et al., 2010). Indeed, the fact that the production effect appears to be primarily restricted to within-subject designs has been taken as evidence for the role of relative distinctiveness as the underlying mechanism. Nevertheless, Fawcett (2013) did report a numerically small but significant between-subjects production effect for reading aloud, prompting alternative accounts to distinctiveness (for review, see Ozubko, Gopie, & MacLeod, 2012, as well as Bodner & Taikh, 2012, and Taikh & Bodner, 2016). Although there is certainly value in seeking to understand how production effects could be produced in a between-subjects design, we remain convinced that distinctiveness is likely the *primary* mechanism that gives rise to production effects – for singing as well as for other forms of production reported in the literature. It seems likely that in between-subjects designs participants may simply be *less* likely to use the distinct information/elements as retrieval cues at test when there was no backdrop of non-produced items during study. In other words, the use of a distinctiveness heuristic is probabilistic rather than absolute.

In the absence of compelling evidence to the contrary, our conclusion is that relative distinctiveness is the most parsimonious account of how various vocal and non-vocal forms of production result in improved recognition performance relative to reading silently. Singing is a more effective production than reading aloud. But the

underlying mechanism giving rise to the production effect is the same in both cases: Relative distinctiveness accounts for better recognition of items read aloud compared to items read silently and also for better recognition of items sung compared to items read aloud.

Acknowledgements

We thank Carl Helmick for designing the custom software used to randomize the items and our undergraduate participants for volunteering their time to contribute to this research. We also thank Ian Palmer for his time and effort in collecting the data for Experiment 2 and Kathy Otten for her time and effort in collecting the data for Experiment 3. Experiments 1, 3, and 4 were first reported in a dissertation submitted by CKQ in partial fulfillment of the requirements for a PhD from Dalhousie University. Support for this study was provided by the Natural Sciences Engineering and Research Council of Canada (NSERC) through a Vanier Scholarship to CKQ and a Discovery Grant to TLT.

References

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.

Bodner, G. E., & MacLeod, C. M. (2016). The benefits of studying by production… and of studying production: Introduction to the special issue on the production effect in memory. *Canadian Journal of Experimental Psychology, 70*, 89-92.

Castel, A. D. (November, 2009). *Predicting the production effect: The use and misuse of self-generated cues*. Paper presented at the 5th bi-annual meeting of the International Association on Metacognition, Boston, MA.

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 judgments of learning. *Memory & Cognition*, *41*, 28-35.

Cohen, J. D., MacWhinney, B., Flatt, M., & Provost, J. (1993). PsyScope: An interactive graphic system for designing and controlling experiments in the psychology laboratory using Macintosh computers. *Behavior Research Methods, Instruments, and Computers, 25*, 257–271.

Cooper, E. C., & Pantle, A. J. (1967). The total-time hypothesis in verbal learning. *Psychological Bulletin, 68*, 221-234.

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 Bulletic & Review*, *8*, 155-161.

Engelkamp, J., Zimmer, H. D., & Biegelmann, U. E. (1993). Bizarreness effects in verbal tasks and subject- performed tasks. *European Journal of Cognitive Psychology*, *5*, 393-415.

Engelkamp, J., Zimmer, H. D., Mohr, G., & Sellen, O. (1994). Memory of self-performed tasks: Self- performing during recognition. *Memory & Cognition*, *22*, 34-39.

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

Fawcett, J. M., & Ozubko, J. D. (2016). Familiarity, but not recollection, supports the between-subject production effect in recognition memory. *Canadian Journal of Experimental Psychology, 70*, 99-115.

Fawcett, J. M., Quinlan, C. K., & Taylor T. L. (2012). Interplay of the production and picture superiority effect: A signal detection analysis. *Memory*, *20*, 655-666.

Forrin, N. D., & MacLeod, C. M. (2016). Auditory presentation at test does not diminish the production effect in recognition. *Canadian Journal of Experimental Psychology, 70*, 116-124.

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

Hassall, C., Quinlan, C. K., Turk, D.J., Taylor, T. L., & Krigolson, O. (2016). A preliminary investigation into the neural basis of the production effect. *Canadian Journal of Experimental Psychology, 70*, 139-146.

Hirshman, E., Whelley, M. M., & Palij, M. (1989). An investigation of paradoxical memory effects. *Journal of Memory and Language, 28,* 594-609.

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

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

Jamieson, R. K., Mewhort, D. J. K., & Hockley, W. E. (2016). A computational account of the production effect: Still playing twenty questions with nature. *Canadian Journal of Experimental Psychology, 70*, 154-164.

Jonker, T. R., Levene, M., & MacLeod, C. M. (2013). Testing the item-order account of design effects using the production effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 40*, 441-448.

MacLeod, C. M., & Bodner, G. E. (2017). The production effect in memory. *Current Directions in Psychological Science, 26,* 390-395.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.

Macmillan, N., & Creelman, C. D. (2005). *Detection theory: A user's guide*. Mahwah, NJ: Lawrence Erlbaum.

Mather, M., & Carstensen, L. L., (2005). Aging and motivated cognition: The positivity effect in attention and memory. *Trends in Cognitive Sciences, 9*, 496-502.

McDaniel, M. A., Anderson, D. C., Einstein, G. O., & O'Halloran, C. M. (1989). Modulation of environmental reinstatement effects through encoding strategies. *American Journal of Psychology, 102*, 523-548.

McDaniel, M. A., & Bugg, J. M. (2008). Instability in memory phenomena: A common puzzle and a unifying explanation. *Psychonomic Bulletin & Review, 15*, 237-255.

McDaniel, M. A., & Einstein, G. O. (1986). Bizarre imagery as an effective memory aid: The importance of distinctiveness. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 12*, 54-65.

McDaniel, M. A., Einstein, G. O., De Losh, E. L., May, C. P., & Brady, P. (1995). The bizarreness effect: It's not surprising, it's complex. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 21,* 422-435.

Mohr, G., Engelkamp, J., & Zimmer, H. D. (1989). Recall and recognition of self-performed acts. *Psychological Research*, *51*, 18-187.

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

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.

Quinlan, C. K., & Taylor, T. L. (2013). Enhancing the production effect in memory. *Memory*, *21*, 904-915.

R Core Team (2018). *R: A language and environment for statistical computing.* [Computer software]. Retrieved from https://cran.r-project.org/.

Schacter, D. L. Israel, L. & Racine, C. (1999). Suppressing false recognition in younger and older adults: The distinctiveness heuristic. *Journal of Memory and Language*, *40*, 1-24.

Schmidt, S. R. (1991). Can we have a distinctive theory of memory? *Memory & Cognition, 19,* 523-542.

Singmann, H. (2018). *Afex:Analysis of Factorial Experiments*. [R package]. Retrieved from https://cran.r-project.org/package=afex.

Taikh, A., & Bodner, G. E. (2016). Evaluating the basis of the between-group production effect in recognition. *Canadian Journal of Experimental Psychology, 70*, 186-194.

The jamovi project (2019). *Jamovi*. (Version 0.9) [Computer Software]. Retrieved from https://www.jamovi.org.

Worthen, J. B., & Wood, V. V. (2001). Memory discrimination for self-performed and imagined acts: Bizarreness effects in false recognition. *The Quarterly Journal of Experimental Psychology, 54A*, 49-67.

Table Captions

Table 1

Means (and standard deviations) for the proportion of hits and false alarms for each experiment (Experiment 1, Experiment 2, Experiment 3, Experiment 4) as a function of production condition (Sing, Read Aloud, Read Silently).

		Hits	Foil False Alarms	Corrected Hits
Experiment 1	Sing	.591 (.128)	.137 (.102)	.454 (.138)
	Read Aloud	.484 (.122)	.121 (.084)	.362 (.137)
	Read Silently	.353 (.133)	.119 (.095)	.234 (.090)
Experiment 2	Sing	.699 (.156)	.213 (.136)	.486 (.180)
	Read Aloud	.607 (.165)	.194 (.116)	.412 (.194)
	Read Silently	.503 (.176)	.186 (.135)	.317 (.168)
Experiment 3	Sing Quickly	.621 (.128)	.142 (.109)	.479 (.143)
	Read Aloud Slowly	.563 (.186)	.150 (.119)	.413 (.191)
	Read Silently	.374 (.173)	.162 (.124)	.212 (.127)
Experiment 4	Sing	.617 (.160)	.192 (.144)	.425 (.139)
	Read Aloud	.581 (.115)	.124 (.079)	.457 (.110)
	Read Silently	.608 (.146)	.204 (.143)	.403 (.141)