



Singing Does Not Necessarily Improve Memory More Than Reading Aloud

An Empirical and Meta-Analytic Investigation

Jedidiah W. Whitridge¹, Mark J. Huff², Jason D. Ozubko³, Paul C. Bürkner⁴, Chelsea D. Lahey¹, and Jonathan M. Fawcett¹

¹Department of Psychology, Memorial University of Newfoundland, St. John's, NL, Canada

²School of Psychology, The University of Southern Mississippi, Hattiesburg, MS, USA

³Department of Psychology, State University of New York at Geneseo, Geneseo, NY, USA

⁴Department of Statistics, TU Dortmund University, Germany

Abstract: The *production effect* refers to the finding that words read aloud are better remembered than words read silently. This finding is typically attributed to the presence of additional sensorimotor features appended to the memory trace by the act of reading aloud, which are not present for items read silently. Supporting this perspective, the production effect tends to be larger for singing (the *singing superiority effect*) than reading aloud, possibly due to the inclusion of further sensorimotor features (e.g., more pronounced tone). However, the singing superiority effect has not always replicated. Across four experiments, we demonstrate a production effect for items read aloud but observe a singing superiority effect only when items are tested in the same color in which they were studied (with foils randomized to color). A series of meta-analytic models revealed the singing superiority effect to be smaller than previously thought and to emerge only when test items are presented in the same color in which they were studied. This outcome is inconsistent with common distinctiveness-based theoretical accounts.

Keywords: production, memory, singing, distinctiveness



We rely on our memories in nearly all facets of life, from basic survival behaviors to higher learning and spirituality. Unsurprisingly, then, a great deal of research has focused on strategies capable of improving memory for important information, such as generation (e.g., McCurdy et al., 2020; Slamecka & Graf, 1978) or levels of processing (e.g., Craik & Lockhart, 1972). Perhaps one of the simplest strategies is the finding that reading information aloud yields superior memory relative to reading it silently (e.g., Conway & Gathercole, 1987; Gathercole & Conway, 1988; Hopkins & Edwards, 1972), a phenomenon dubbed the *production effect* (MacLeod et al., 2010). This effect has since been shown to be both versatile and robust and to persist across a variety of production modalities (e.g., writing; Forrin et al., 2012; drawing; Wammes et al., 2016), populations (e.g.,

older adults; Lin & MacLeod, 2012; individuals with speech and hearing impairments; Icht et al., 2019; Taitelbaum-Swead et al., 2018), and paradigms (e.g., short- and long-list recall; Cyr et al., 2022; Saint-Aubin et al., 2021).

Since the production effect was first delineated, theorists have sought to identify its underlying cognitive mechanisms (e.g., Gathercole & Conway, 1988; Hopkins & Edwards, 1972). Although these processes remain a subject of debate (e.g., Fawcett, 2013; Fawcett et al., 2023), theoretical perspectives generally contend that the production effect is driven predominantly by encoding distinctiveness (e.g., Dodson & Schacter, 2001; MacLeod et al., 2010). According to this *distinctiveness account*, producing an item encodes additional sensorimotor features (i.e., the production trace; Fawcett, 2013; Fawcett et al., 2012) not present for silent items. At test, participants are thought to use this production trace to guide test performance, either consciously (Dodson & Schacter, 2001) or via unconscious retrieval dynamics (Jamieson et al., 2016). Evidence generally supports this framework – the production effect is eliminated by reducing

the distinctiveness of the productive act (e.g., by using a common vocal response or action; MacLeod et al., 2010; Richler et al., 2013) or obviating the diagnostic value of the production trace (e.g., by having participants produce items from all sources in a list discrimination task; Ozubko & MacLeod, 2010).

One corollary of the distinctiveness account is that the mnemonic benefits afforded by production ought to be positively correlated with the number of distinct sensorimotor features encoded at study (Forrin et al., 2012; also see Fawcett et al., 2012; Jamieson et al., 2016; Kelly et al., 2022; Quinlan & Taylor, 2013). This *sensorimotor scaling hypothesis* has received empirical support. For example, the production effect is larger for reading aloud than it is for writing (e.g., Forrin et al., 2012) or mouthing (e.g., Gathercole & Conway, 1988). Whereas reading aloud incorporates visual, motoric, and auditory features, writing and mouthing exclude the auditory component (also see Fernandes et al., 2018). Furthermore, when words are presented auditorily (rather than visually), the production effect becomes smaller for items read aloud compared to those written, possibly owing to the elimination of visual features in the former case (Mama & Icht, 2016).

Another piece of evidence favoring the sensorimotor scaling hypothesis is the finding that the incorporation of tonal and rhythmic information – via singing – produces an especially large production effect that exceeds reading aloud (Quinlan & Taylor, 2013, 2019). Quinlan and Taylor (2019) subsequently replicated and extended this *singing superiority effect* (SSE), ruling out alternative explanations, such as bizarreness, differential production speed, and differences in the strength of encoding. They argued that singing results in an especially elaborate production trace. The SSE has since been accepted as evidence of both the sensorimotor scaling hypothesis and the distinctiveness account (e.g., Forrin & MacLeod, 2018; Mama & Icht, 2016).

However, there have also been several failed replications of the SSE (Hassall et al., 2016; also see Ozubko et al., 2020). Furthermore, additional efforts to increase the magnitude of the production effect by inducing especially distinctive forms of production (e.g., character voices; Wakeham-Lewis et al., 2022) have similarly failed to produce a memory boost (or even eliminated the benefit altogether). Moreover, only three published studies have investigated the production effect for singing and most of those experiments used small samples (i.e., $Ns < 24$ participants) and produced variable effect sizes (i.e., $d = \sim 0.3 - \sim 1.5$; Quinlan & Taylor, 2013, 2019). Given that support for the notion that additional distinctive features can increase the magnitude of the production effect rests almost entirely upon the SSE, the reliability of this effect is critically important to modern theoretical frameworks.

To address this, we conducted four experiments replicating the SSE. Experiments 1a and 1b conceptually replicated Quinlan and Taylor (2013; Experiment 2) with the exceptions

that we assessed recollection and familiarity (Yonelinas, 2002) using recollect/familiar/neither judgments and did not present test items in the colors in which they had been studied. Experiment 2 modified our approach to incorporate color-matched target and foil items at test as used by Quinlan and Taylor (2013; see Fawcett et al., 2012, for a detailed discussion). This change was undertaken to investigate the possibility that orienting participants to study conditions via stimulus dimensions might lead to the use of atypical retrieval strategies at test. Experiment 3 incorporated further methodological changes to replicate Quinlan and Taylor (2013; Experiment 2) as closely as possible. Finally, Experiment 4 investigated whether the SSE would replicate in a between-subject design, as this had not been shown in the past (e.g., Quinlan & Taylor, 2019; Experiment 4). Additionally, we report a meta-analysis of all known studies of the SSE. Experiments 1b, 2, and 4 were pre-registered. Pre-registrations for these experiments are available on the Open Science Framework website (<https://osf.io/z6jue>).

To preview, we observed a robust production effect for singing and reading aloud across designs. However, a credible SSE only emerged for the color-matched group in Experiment 3, hinting at the possibility that knowing the study condition of a given test item may facilitate the effect. Our meta-analytic model estimated a small but credible aggregate SSE, although moderator analysis revealed this to be driven by studies using the color-matching procedure.

Experiments 1a and 1b

In both Experiments 1a and 1b, production (sing, aloud, silent) was manipulated within-subjects with a confidence-based recognition response at test, followed by recollect/familiar/neither judgments. The latter was included to evaluate whether the production effect for singing was driven by episodic-based recollective processes, fluency-based familiarity processes, or a combination of both. Because these studies were similar and produced near identical results, their data were combined for analysis. Further details and analyses of the individual experiments are provided in Electronic Supplementary Material 1 (ESM 1). On the basis of recent research into interactions between the production effect and serial position (e.g., Gionet et al., 2024; Saint-Aubin et al., 2021), we additionally conducted exploratory serial position analyses of the present experiment and all experiments reported hereafter; these analyses are reported in ESM1.

Method

Participants

Experiment 1a consisted of 25 undergraduates as participants from the State University of New York at Geneseo,

and Experiment 1b consisted of 43 undergraduates as participants from The University of Southern Mississippi, who participated for partial course credit. One participant from Experiment 1a was excluded from analyses due to accuracy far below chance (false alarm rate > hit rate across all conditions). Details pertaining to sample size determination and exclusions for all experiments can be found in our pre-registration (<https://osf.io/z6jue>).

Stimuli and Apparatus

Stimuli for Experiment 1a and 1b consisted of 348 and 360 words, respectively; details about the stimulus sets are available in ESM 1. In both experiments, participants were randomly assigned a subset of 180 words. Half were randomized between the three study conditions (30 silent, aloud, and sing). Words at study were presented in red, yellow, or blue font indicating study condition; color assignments were counterbalanced. The remainder of the words appeared as “new” foils at test and were presented in white font. Experiments 1a and 1b were coded in PsychoPy (versions 1.84.2 and 2.3.2, respectively; Peirce et al., 2019) and presented via a color monitor (17-in. and 20-in., respectively) attached to a computer running Windows (versions 8 and 10, respectively). For both experiments, all stimuli were presented in 14-point Arial font against a black background.

Procedure

Experiments 1a and 1b were identical except where specifically noted. Each experiment consisted of a study phase and a test phase. In Experiment 1a, participants were simply instructed to read the words silently, aloud, or by singing (depending on color). In Experiment 1b, however, the experimenter provided a demonstration to each participant for how words should be produced, with emphasis on the singing condition. Participants in that experiment were further instructed to sing as effortfully as they could and to differentiate their singing tonally from their typical reading voice.

Study Phase

During the study phase, participants were presented with 90 words, one at a time, with one-third in each production condition; words from the three conditions were intermixed and presented in random order. Each trial began with a 500-ms fixation (“+”), followed by a 500-ms blank screen and then the word at center for 2000 ms. An experimenter remained present throughout the study phase. In Experiment 1b, the experimenter monitored study responses to ensure participants were singing in a manner that adequately distinguished that condition from reading aloud. If the experimenter deemed that the participant was not singing adequately, they were encouraged to sing with greater gusto. Following presentation of all study items, participants proceeded to the test phase.

Test Phase

During the test phase, participants were presented with a total of 90 “old” and 90 new words, each in white font. Test trials began with a 500-ms fixation “+”, followed by a 500-ms blank screen and the word at center. The word remained on screen until participants made both a confidence judgment and a recollect/familiar/neither judgment, which were separated by a 500-ms blank screen.

Confidence judgments were given as a rating on a scale ranging from 1 to 6. Values from 1 to 3 indicated that participants thought the word was new, whereas values from 4 to 6 indicated confidence that the word was old. Anchors were provided for each value: Confidence in the new or old status of the word could be *less sure*, *somewhat sure*, or *very sure*, with values of 1 or 6 indicating maximum confidence that the word was new or old, respectively. The recollect/familiar/neither judgment was analogous to commonly employed remember/know/no judgments (e.g., Fawcett & Ozubko, 2016). Responses were given by pressing the “R” key to indicate the word was *recollected* (i.e., remembered), the “F” key to indicate that the word was *familiar* (i.e., known), or “N” to indicate that the word was neither recollected nor familiar.

Statistical Approach

Our approach utilized Bayesian probit regression models to estimate metrics similar to d' and C in a multilevel context. We view this approach as advantageous for several reasons. First, our primary dependent measures were binary. While binary outcome data (e.g., old or new responses) are often aggregated into proportions, this procedure violates assumptions made by typical statistical approaches (e.g., Baayen et al., 2002; Jaeger, 2008). Furthermore, previous studies have argued for the superiority of signal detection analysis over raw hits for interpretation of the production effect (e.g., Fawcett et al., 2012, 2023). Similarly, the present experiments aimed to evaluate evidence for or against the existence of an effect. Bayesian approaches allow for the quantification of evidence favoring a null model (Masson, 2011). For these reasons, we utilized multilevel probit regression models implemented via the *brms* package (Bürkner, 2017) in R (R Core Team, 2020). For interested readers, detailed information about our models (including random effect structures, priors, and model fit parameters) for the present experiment and all models reported hereafter is provided in ESM 1 (also see Fawcett & Ozubko, 2016; Fawcett et al., 2016).

Results and Discussion

Confidence Ratings

Confidence ratings were binarized such that ratings greater than three indicated an old response. We then

applied a multilevel probit regression to the binary responses with item type (sing, aloud, silent, foil) as a fixed effect. Because Experiments 1a and 1b did not separate false alarm rates by condition, it was not meaningful to interpret C .

For each model, we report median posterior estimates for d' by condition and for contrasts between conditions. The latter parameters were calculated directly from the posterior distributions of the estimates for each condition and reflect raw differences in d' . Alongside these parameters, we report the 95% highest density interval (HDI) surrounding each estimate. The HDI represents the interval containing 95% of the posterior distribution such that all values within the interval are more probable than values that fall outside the interval (Kruschke, 2010). Intervals excluding 0 suggest that 0 is not a credible value, analogous to statistical significance in frequentist models.

As shown in Figure 1, we observed a credible production effect for both reading aloud and singing. However, we did

not observe a credible SSE (difference = 0.00, $\text{HDI}_{95\%} = -0.13$ to 0.12). These findings pose an initial challenge to the reliability of the SSE.

Recollection

Having evaluated the SSE in standard recognition, we next applied a comparable multilevel probit model to analyze “recollect” responses. Recollection is often viewed as a measure of episodic memory or re-experiencing (Yonelinas, 2002). Analyzing recollection responses produces estimates analogous to d' , only reflecting the degree to which participants differentiated between new and old items via their recollect responses.

As shown in Figure 1, we observed production effects within recollection for both singing and reading aloud that were of similar magnitude to the production effects observed for confidence ratings. There was no credible difference between the sing and aloud conditions (difference = -0.01 , $\text{HDI}_{95\%} = -0.13$ to 0.12). As expected,

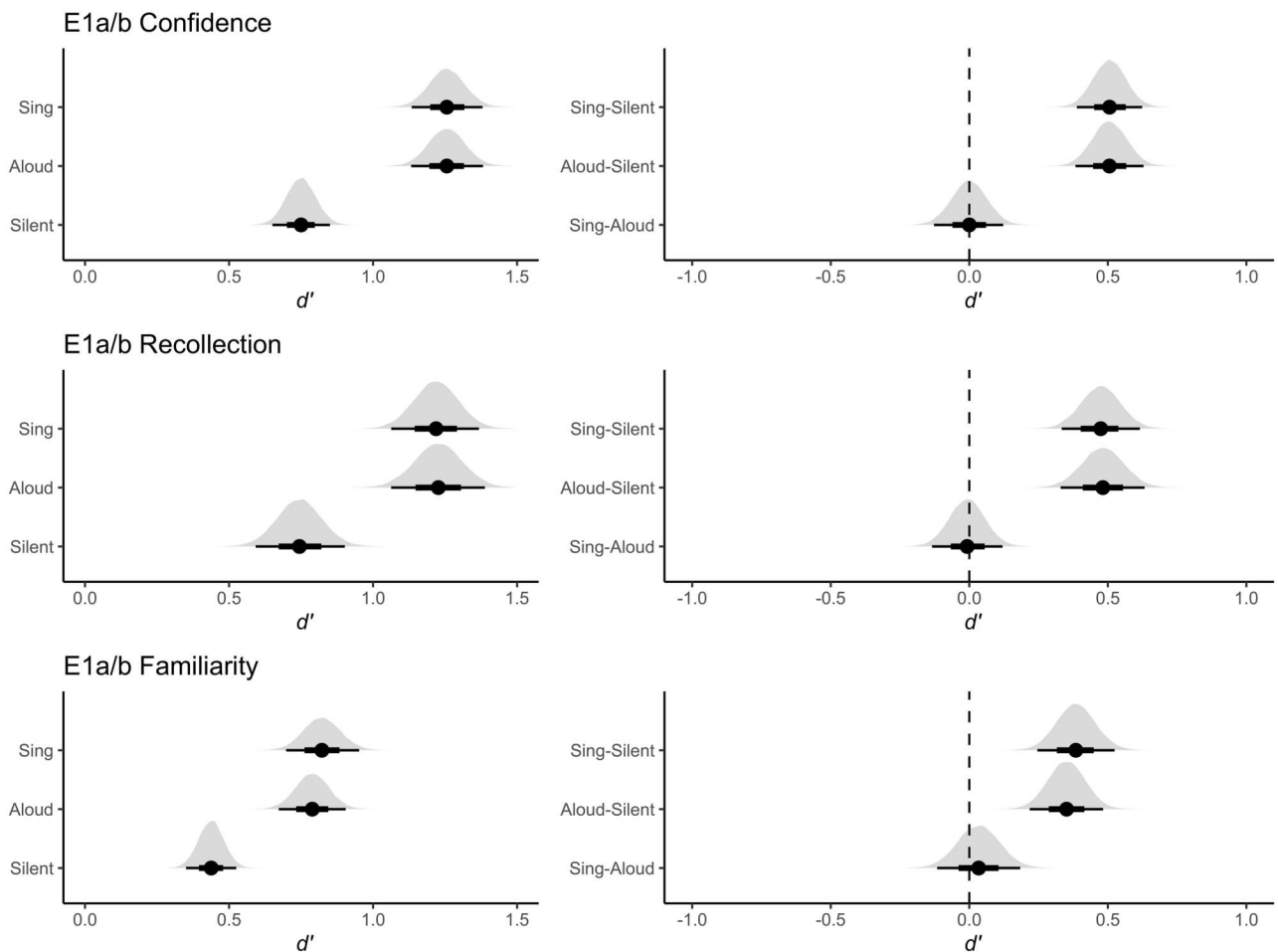


Figure 1. Posterior estimates for sensitivity (d') as a function of condition (left column) and contrasts between conditions (right column) for Experiment 1. Note. Polygons depict the posterior distribution for each estimate, and points show the median estimate. Thick lines represent the 50% HDI, and thin lines represent the 95% HDI.

this model replicates earlier research showing that the within-subject production effect for reading aloud is driven in part by recollective processes (e.g., Fawcett & Ozubko, 2016; Ozubko et al., 2012) and extends this finding to the production effect for singing.

Familiarity

Finally, we analyzed familiarity. Familiarity is often viewed as a nonspecific feeling of fluency or familiarity that can drive recognition responses (Yonelinas, 2002). Familiarity was analyzed using an analog of the Independence Remember-Know (IRK; e.g., Yonelinas, 2002; Yonelinas & Jacoby, 1995) procedure. Specifically, we applied a multilevel probit model analogous to those used above (albeit predicting familiar responses) to trials for which recollect responses were not made. This approach is equivalent to conventional calculations of the IRK procedure (for further discussion and mathematical proof, see Fawcett et al., 2016; also see Fawcett & Ozubko, 2016).

As shown in Figure 1, the production effect for familiarity was credible for both reading aloud and singing, but with little difference between the two critical conditions (difference = 0.03, $HDI_{95\%} = -0.12$ to 0.18). Replicating earlier work (Fawcett & Ozubko, 2016; Ozubko et al., 2012), it appears that the production effect for singing is driven by both recollection and familiarity in within-subject designs.¹

Experiment 2

Having failed to replicate the SSE, we considered the possibility that the effect might be driven by some unknown methodological factor. Specifically, Experiments 1a and 1b deviated from the methods used by Quinlan and Taylor (2013). One difference was Quinlan and Taylor's use of color matching at test, a procedure first used by Fawcett et al. (2012) to permit separate false alarm rates and thereby the calculation of C and d' for each condition. We initially opted *not* to use this procedure for two reasons. First, distinctiveness accounts of the production effect predict that the SSE should arise on the basis of additional sensorimotor features appended to the production trace (Forrin et al., 2012; Quinlan & Taylor, 2013, 2019), which should be agnostic to whether participants are aware of the study condition for a specific test item. Second, presenting words at test in their study phase colors introduces context

effects, which are known to impact memory (e.g., Isarida & Isarida, 2007; Tulving & Thomson, 1973).

However, we speculated that this decision might have obfuscated the SSE. Some variants of the distinctiveness account contend that participants leverage knowledge of having produced information consciously in the form of a *distinctiveness heuristic* to guide discrimination (i.e., "I remember saying it aloud, so I must have studied it"; Dodson & Schacter, 2001; also see MacLeod et al., 2010). By orienting participants to stimulus dimensions via item color, it is possible that participants might deploy heuristics for each condition. For example, recognizing that test items presented in red were sung at study might encourage participants to monitor for tonal information, which may have been missed if the study condition was not cued. To explore this possibility, Experiment 2 replicated the procedure used in Experiments 1a and 1b, albeit with the inclusion of test-phase color matching as a between-subject manipulation.

Method

Participants

Experiment 2 consisted of 90 undergraduates ($N = 45$ matched) as participants from The University of Southern Mississippi who completed the experiment in exchange for partial course credit.

Materials and Procedure

Experiment 2 was identical to Experiment 1b with the exception that the matched group received words at test in their corresponding study phase colors, with foil items split equally across colors (30 each).

Statistical Approach

The statistical approach taken for Experiment 2 was similar to Experiments 1a and 1b in that we estimated d' and C using multilevel probit regression models. However, because we recorded separate false alarm rates for each condition in this experiment (foils were arbitrarily assigned to production condition within the nonmatched condition), the parameterization of the models differed slightly. The availability of separate false alarm rates allowed d' and C to be modeled using a nonlinear equation with fixed effects for condition (sing, aloud, silent) and group (matched, unmatched) applied separately to each parameter.

¹ Consistent with the results we reported here, an unpublished investigation by Zhang (2024) observed robust production effects for both reading aloud and singing on recollection and IRK familiarity hit rates. In this case, the authors did not observe an SSE in recollection and in fact observed a small advantage for reading aloud (i.e., aloud > sing) in familiarity.

Results and Discussion

For all dependent measures, we report d' in text. For the present experiment and all experiments reported hereafter, analyses of C are available in ESM 1.

Confidence Ratings

As depicted in Figure 2, the production effects for either modality were credible across groups, although the SSE failed to credibly emerge in either matching condition (with a slight positive trend in the matched and a slight negative trend in the unmatched conditions). A numerical trend also favored a larger SSE in the matched group (difference = 0.15, $HDI_{95\%} = -0.16$ to 0.46), hinting at a

potential interaction between singing and color matching. We return to this point in Experiment 3.

Recollection

As shown in Figure 2, the production effects within recollection for either modality were credible across groups, with similar numeric trends in the matched group as observed for the confidence ratings (difference = 0.18, $HDI_{95\%} = -0.18$ to 0.53).

Familiarity

As also depicted in Figure 2, analysis of the familiarity responses followed the same pattern observed for our other dependent measures, with production effects for either

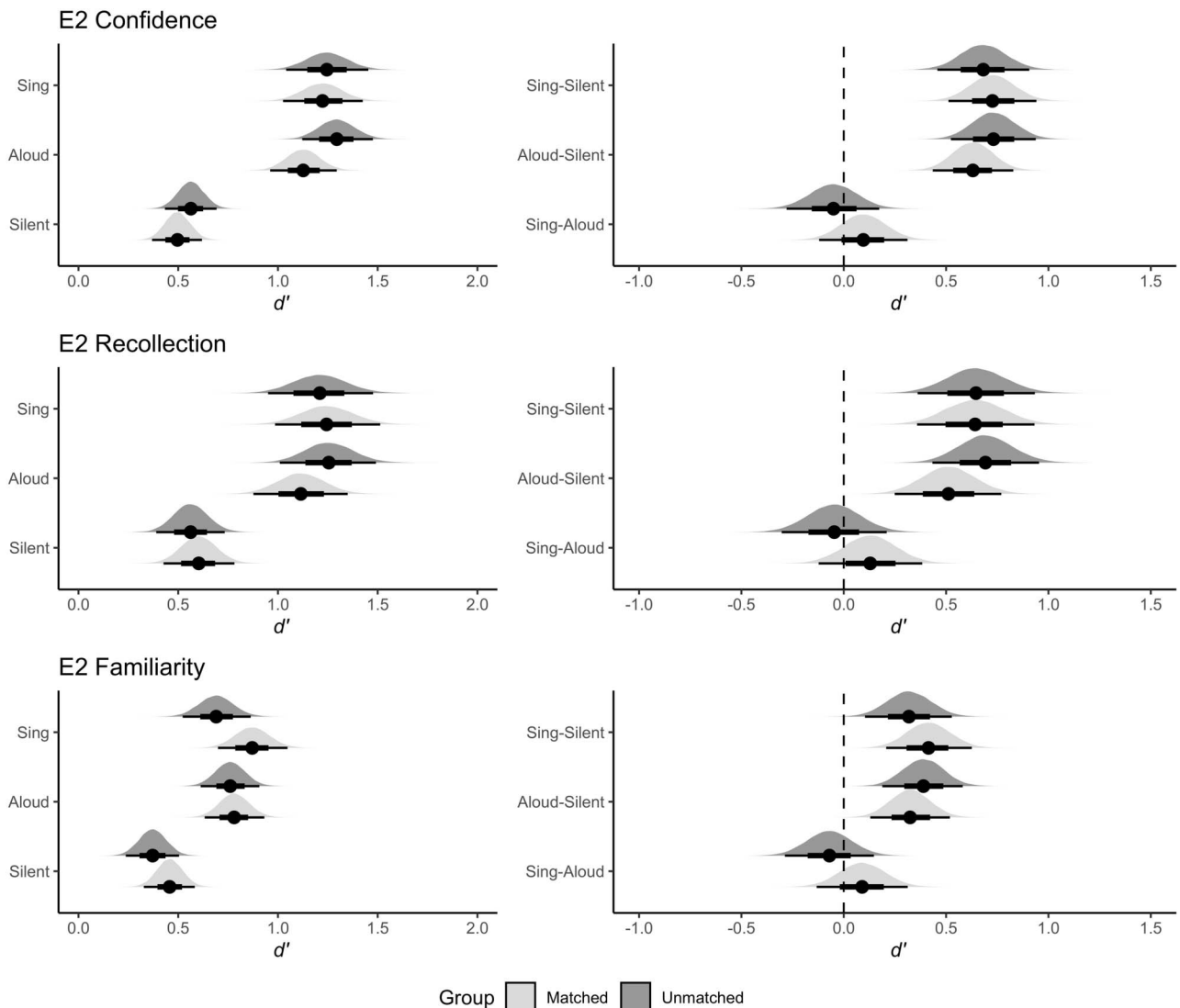


Figure 2. Posterior estimates for sensitivity (d') as a function of condition and group (left column) and contrasts between conditions as a function of group (right column) for Experiment 2. Note. Polygons depict the posterior distribution for each estimate, and points show the median estimate. Thick lines represent the 50% HDI, and thin lines represent the 95% HDI.

modality and little support for an SSE or a difference in the magnitude of the SSE between matching conditions (difference = 0.16, $HDI_{95\%} = -0.15$ to 0.47). Overall, Experiment 2 replicated Experiments 1a and 1b but again failed to detect a credible SSE across any dependent measure.

Experiment 3

Although not credible, the trends observed in Experiment 2 suggested that the SSE may be slightly larger when study condition is provided at test via color matching. Furthermore, our methodology deviated from earlier demonstrations of the SSE in other unexplored ways. Thus, the purpose of Experiment 3 was to rule out other potential hidden moderators by directly replicating Quinlan and Taylor (2013, Experiment 2). To this end, we added pre-study phases and modified stimuli, timings, and the recognition test to match that experiment.

Method

Participants

Experiment 3 consisted of 102 undergraduates ($N = 51$ matched) as participants from The University of Southern Mississippi who completed the experiment in exchange for partial course credit.

Stimuli and Apparatus

Stimuli in this experiment consisted of 240 words; details about this stimulus set are reported in ESM 1. Each participant saw all possible words over the course of the experiment. Half the words appeared in the study phase and were randomized between the three study conditions for each. Words at study were presented in colored font, with each respective study condition being assigned red, white, or blue. Color assignments were counterbalanced across participants. The remainder of the words appeared only as new foils at test. For the matched group, the color assignment for foils was randomized across the three possible assignments. For the unmatched group, all foils were presented in yellow. All words were presented in 42-point Times New Roman font against a black background. The apparatus was identical to Experiment 1b.

Procedure

The experiment consisted of a familiarization, practice, study, and test phase. Prior to familiarization, participants were informed that they would see words presented one at a time in one of three colors (red, white, or blue), which indicated how the words should be studied. Instructions

for each condition were derived directly from Quinlan and Taylor (2013). Participants were told that they would complete a memory test after they had studied all the words. The experimenter remained in the room with participants throughout the familiarization phase, practice phase, and study phase.

Familiarization Phase

In the familiarization phase, participants were presented with 15 trials. Participants saw five familiarization trials per study condition (i.e., sing, read aloud, read silently) in random order. Each trial consisted of a 500-ms blank screen followed by the name of a color assignment and its associated study condition (e.g., RED - Sing) for 2000 ms; text in each familiarization trial was displayed in colored font corresponding to the color assignment being displayed. After all familiarization trials had been presented, participants moved on to the practice phase.

Practice Phase

The practice phase consisted of 15 trials, five per study condition, presented in random order. Each trial consisted of a 500-ms blank screen followed by the presentation of the word “banana” at center and in colored font for 2000 ms. As indicated by the word’s color assignment, participants were cued to either sing the word, read it aloud, or read it silently. The study phase immediately followed.

Study Phase

The study phase was identical to Experiment 2 with two exceptions. First, participants were presented with a total of 120 words. Second, study trials consisted of a 500-ms blank screen followed by the word at center for 2,000 ms. After all study trials were complete, participants moved on to the test phase.

Test Phase

During the test phase, participants were presented with a total of 240 words. The color assignment of test words was similar to Experiment 2, albeit with the exceptions that 40 (rather than 30) foil items were presented in each possible color (i.e., red, white, or blue) for the matched group and that all words were presented in yellow for the unmatched group. There were no other differences between the matched and unmatched groups. Each test trial began with a 500-ms blank screen followed by the word at center. The word remained on screen until participants made a yes/no recognition judgment. Judgments were made using a textbox that appeared below the word, in which participants could respond to the word by pressing either the “Y” key (yes) or the “N” key (no). Participants could correct their responses using the backspace key. When ready, participants submitted the response to each trial using the

space bar. After each response, the next word was presented at center; this procedure repeated until participants had completed all 240 trials.

Statistical Approach

For Experiment 3, all models were as described earlier, albeit applied to *yes* and *no* responses.

Results and Discussion

As depicted in Figure 3, whereas both modalities produced credible production effects, the SSE was credible only in aggregate or in the matched condition (difference = 0.23, $HDI_{95\%} = 0.03$ to 0.43). These results suggest that the SSE is driven in part by the foil matching procedure, although we failed to detect a credible difference related to group, with the overall effect of this parameter estimated at 0.14 ($HDI_{95\%} = -0.02$ to 0.31).

Experiment 4

Prior to fully exploring the observed SSE, we first sought to address the claim that the production effect for singing does not occur in between-subject designs (Quinlan & Taylor, 2019; Experiment 4). Because between-subject

designs remove the *backdrop* against which produced items can be distinctive relative to unproduced items (MacLeod et al., 2010), previous failures to observe a production effect for such designs were viewed as evidence that any benefit of singing must be driven by relative distinctiveness, similar to early perspectives on the benefits of reading aloud. Although this framework was supported by early research (e.g., Conway & Gathercole, 1987; Dodson & Schacter, 2001; Hopkins & Edwards, 1972; MacLeod et al., 2010), a great deal of evidence now supports the notion that the production effect is robust in between-subject designs (e.g., Bodner et al., 2014, 2016; Forrin et al., 2016), albeit smaller than within-subject designs (for meta-analyses, see Fawcett, 2013; Fawcett et al., 2023). Given that Quinlan and Taylor (2019) also failed to observe a production effect for reading aloud, we reasoned that the experiment may have lacked sufficient statistical power. Accordingly, the present experiment tested this hypothesis using a much larger sample size.

Additionally, this experiment is the first to test whether the between-subject production effect for singing is driven by recollection, familiarity, or both. A series of experiments conducted by Fawcett and Ozubko (2016) showed that while the within-subject production effect for reading aloud is driven by both recollection and familiarity, its between-subject counterpart is driven by familiarity alone. Accordingly, the present experiment should provide a further test of the conclusions drawn by Fawcett and Ozubko (2016) and evaluate whether this pattern of results persists for singing.

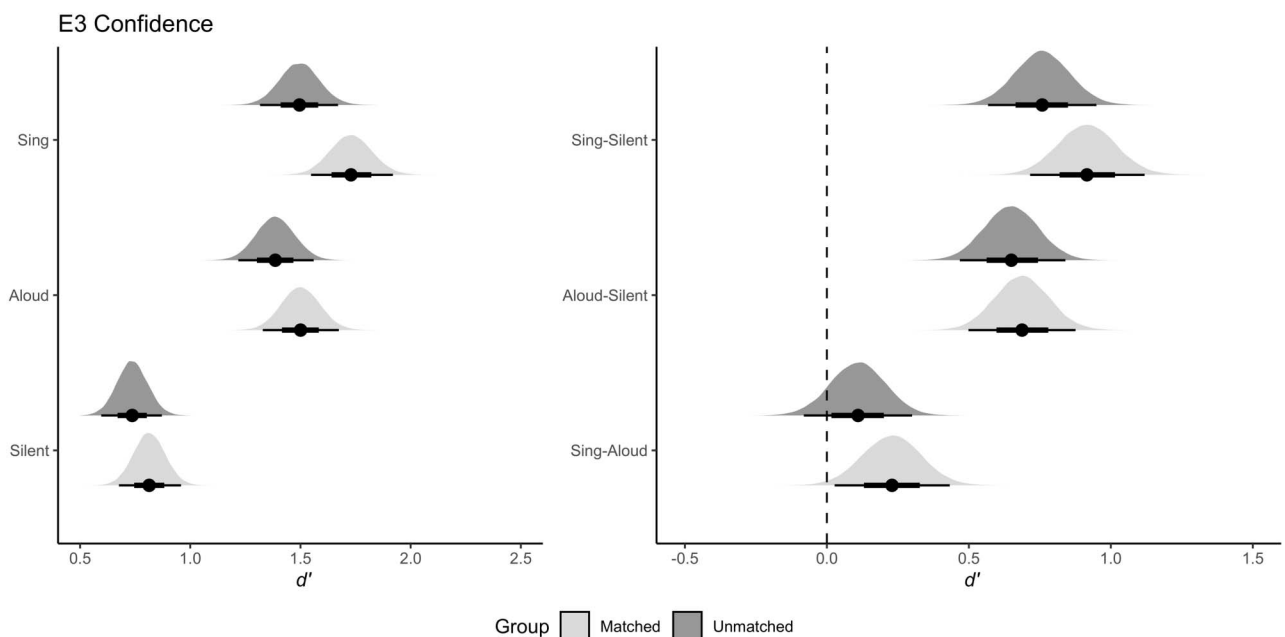


Figure 3. Posterior estimates for sensitivity (d') as a function of condition and group (left column) and contrasts between conditions as a function of group (right column) for Experiment 3. Note. Polygons depict the posterior distribution for each estimate, and points show the median estimate. Thick lines represent the 50% HDI, and thin lines represent the 95% HDI.

Method

Participants

Experiment 4 consisted of 140 undergraduates as participants from The University of Southern Mississippi who completed the experiment in exchange for partial course credit. Fifteen participants were excluded from analyses due to their failure to discriminate between old and new items at above chance level. Participants were randomly assigned to one of three conditions: read silently ($N = 40$), aloud ($N = 42$), or sing ($N = 43$).

Materials and Procedure

The stimuli and apparatus were identical to those described for Experiment 1b. Prior to the experiment, each participant was randomly assigned to one of three production conditions (i.e., sing, aloud, or silent). Prior to the study phase, participants were informed that they would see words presented, one at a time, in one of three colors and that they should ignore the color assignment of each

word. Depending on the condition to which they were assigned, participants were instructed to (1) read all words silently, (2) read all words aloud, or (3) sing all words. The procedure was otherwise identical to Experiment 1b.

Statistical Approach

The statistical approach taken for Experiment 4 was identical to that described for Experiment 2, except for the fact that models for this experiment included only a fixed effect for condition. Because condition was manipulated between-subjects, the random effect structure of the models also differed; details can be found in ESM 1.

Results and Discussion

Confidence Ratings

As shown in Figure 4, a credible production effect was observed for both singing and reading aloud. However, a noncredible numerical trend favored higher sensitivity for

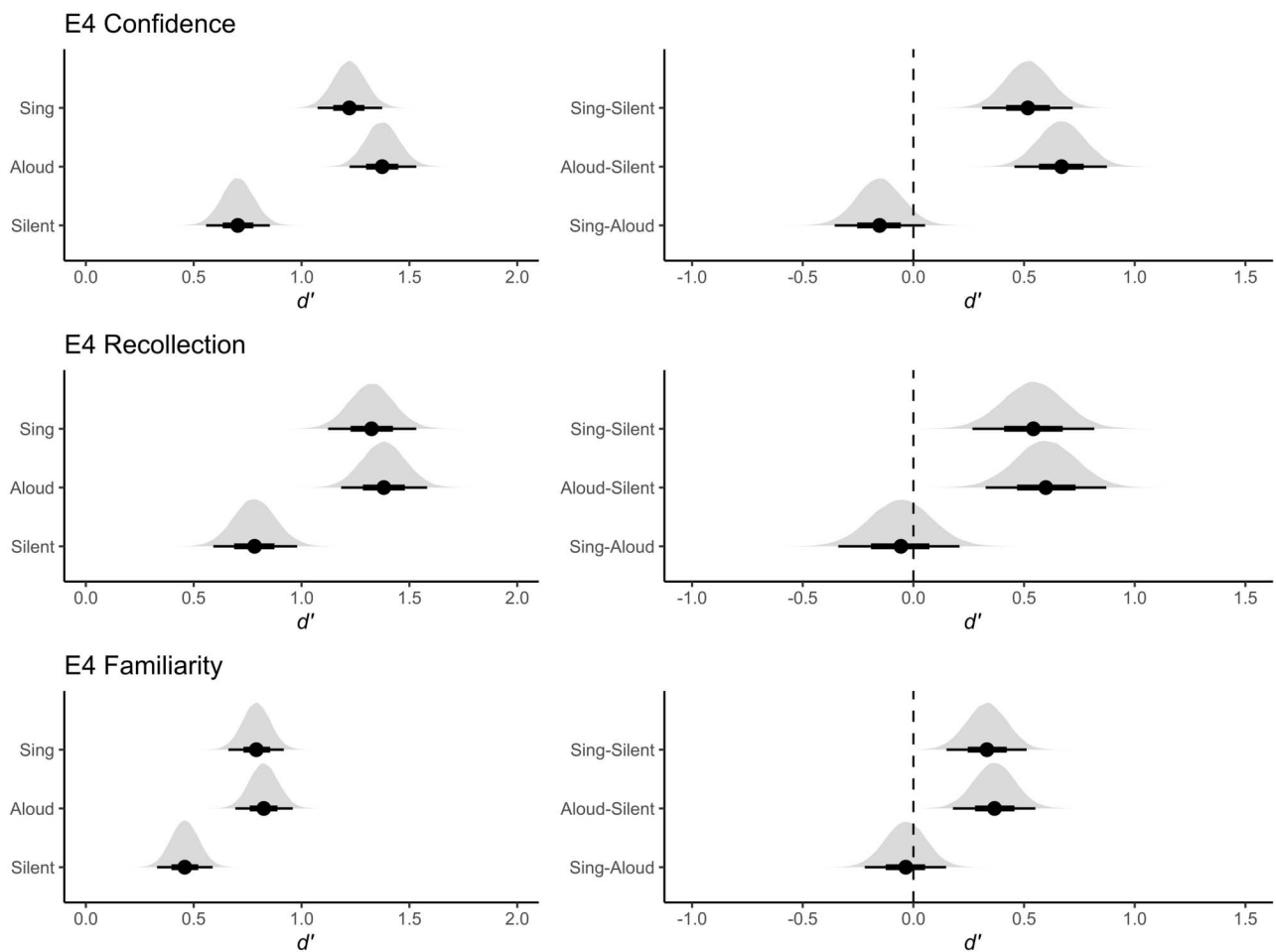


Figure 4. Posterior estimates for sensitivity (d') as a function of condition (left column) and contrasts between conditions (right column) for Experiment 4. Note. Polygons depict the posterior distribution for each estimate, and points show the median estimate. Thick lines represent the 50% HDI, and thin lines represent the 95% HDI.

the aloud relative to the sing condition (difference = -0.15 , $\text{HDI}_{95\%} = -0.36$ to 0.05). Interestingly, the production effects we observed for both singing and reading aloud were comparable in magnitude to our earlier within-subject experiments. This contrasts with prior evidence, suggesting that the between-subject production effect is typically *smaller* than its within-subject counterpart (e.g., Bodner et al., 2014; Fawcett, 2013; Fawcett et al., 2023). This unusual pattern is discussed below.

Recollection

As depicted in Figure 4, production effects were observed for either production modality, but again, there was no evidence of an SSE (difference = -0.06 , $\text{HDI}_{95\%} = -0.34$ to 0.21). However, the presence of a production effect for recollection in a between-subject design fails to replicate the findings of Fawcett and Ozubko (2016).

Given that the emergence of a between-subject production effect for recollection coincides with an unusually large production effect for this design, we speculate that some unknown aspect of our methods caused the recollective component to reappear. Although no candidate moderators are apparent, our study differed from Fawcett and Ozubko (2016) in that participants were supervised by a researcher for the entirety of the study phase. While minor, it is possible that supervision at study may have encouraged participants to remain attentive, leading to stronger encoding and thereby more detailed item representations consistent with a recollective experience. Consistent with this possibility, Bodner et al. (2016) tested participants in small groups and observed within- and between-subject production effects of comparable magnitude. While recollection was not assessed, these findings are congruent with the notion that participants might pay more attention to production tasks in the presence of others. Were this the case, however, it is not clear why additional attentional allocation would facilitate memory in the produced conditions preferentially. At present, we cannot satisfactorily account for our observation of between-subject production effects on recollection; further research is necessary to elucidate the mechanisms that might have driven this pattern of results.

Familiarity

As depicted in Figure 4, the production effects for either modality were credible. Much like the analysis of recollection, the difference in familiarity between the sing and aloud conditions was centered on zero, estimated at -0.03 ($\text{HDI}_{95\%} = -0.22$ to 0.15). These findings align well with earlier research: Like the production effect for reading aloud, the benefit for singing is driven – at least partially – by an increase to familiarity in between-subject designs.

Meta-Analysis of the Singing Superiority Effect

Having largely failed to replicate the SSE (with the exception of Experiment 3), we finally opted to conduct a meta-analysis of the extant literature on this topic to provide a stronger empirical test of this phenomenon. Details about our search and coding procedures can be found in ESM 1.

Method

Effect Size Calculation and Statistical Approach

For all models, effect sizes were calculated as raw difference scores computed using the *escalc* function from the *metafor* package (Viechtbauer, 2010) in *R* (R Core Team, 2020). As our primary dependent measure across experiments has been sensitivity (rather than raw or corrected hits), we computed effect sizes for each experiment as the raw mean difference in d' scores between relevant conditions. Raw data were procured for all studies with the exception of Experiment 3 from Quinlan and Taylor (2013), for which mean d' scores for each condition were coded directly from the article. For all studies for which raw data could be obtained, d' was calculated by aggregating hits and false alarms into proportions and applying transformations to the data (see, e.g., Stanislaw & Todorov, 1999). Because estimates of variability for differences between conditions were not available for Quinlan and Taylor (2013, Experiment 3), we imputed this parameter, as is standard practice.

Models were fit using the *brms* package (Bürkner, 2017) in *R* (R Core Team, 2020) using an approach comparable to frequentist random-effects meta-analysis. We opted to use a Bayesian approach for two reasons. First, simulation studies show that Bayesian models provide superior estimates of both aggregate effects and between-study heterogeneity, particularly with few effects (e.g., Williams et al., 2018). Second, Bayesian models produce credible intervals that allow for probabilistic statements to be made regarding the existence of effects in the data, permitting direct and intuitive interpretation of effects (e.g., Morey et al., 2016).

We modeled our data in two different ways – first including the SSE (sing-aloud) as our effect of interest and a separate model including the production effect (sing/aloud-silent) observed in our singing and read aloud conditions with production modality as a moderator. The parameterization of these models differed slightly. For models of the SSE, we included random effects corresponding to each unique effect size. For models of the production effect, we included a fixed effect for production modality (sing,

aloud) and random effects corresponding to each sample from which the effects were derived. Because the latter models used a common comparison condition for each effect (i.e., silent), our approach allowed for the estimation of separate dependent effects for each sample.

Results and Discussion

For each model, we report median posterior estimates reflecting the raw mean difference in d' for each relevant comparison alongside the 95% HDI. Where applicable, we also report 95% prediction intervals (PIs), which reflect the range of plausible *true* effects expected from hypothetical studies similar to those included in our sample (IntHout et al., 2016).

Models of the Singing Superiority Effect

As depicted in Figure 5, the aggregate SSE was credible, with the difference between the sing and aloud conditions estimated at 0.13 (HDI_{95%} = 0.03 to 0.24). Although the estimate is credible, this model also implies that the size of the effect is much smaller than previous experiments have reported (e.g., Quinlan & Taylor, 2013, 2019). Furthermore, this model indicated substantial heterogeneity across effects, with prediction intervals ranging from -0.14 to 0.47; this implies that some studies show unresponsive effects (i.e., aloud > sing), whereas others show effects that are quite large. This pattern of results is

unsurprising given that previous research has often utilized underpowered samples, which are liable to provide poor estimates of the effect due to sampling error (e.g., Wilson Van Voorhis & Morgan, 2007); our model suggests the SSE – if truly reliable – has likely been overestimated.

Given our earlier experiments suggested foil matching might play an important role in facilitating the SSE, we conducted an exploratory analysis that included color matching as a moderator. Here, the aggregate SSE was credible when color matching was present, with the difference between the sing and aloud conditions estimated at 0.22 (HDI_{95%} = 0.11 to 0.33; PI_{95%} = -0.00 to 0.49). However, the effect was not credible in the absence of this procedure, estimated at 0.00 (HDI_{95%} = -0.11 to 0.14; PI_{95%} = -0.22 to 0.29). Consistent with the patterns we observed in Experiments 2 and 3, these results suggest that the SSE might emerge only when color matching is used at test. Publication bias is evaluated in ESM 1; to summarize those findings, although typical measures of publication bias failed to provide strong evidence, the aggregate effect is nonetheless influenced by small studies favoring a large SSE.

Models of the Production Effect

As shown in Figure 6, the aggregate production effects for both aloud and sing conditions were credible, with the differences between the sing/aloud and silent conditions respectively estimated at 0.57 (HDI_{95%} = 0.48–0.66; PI_{95%} = 0.35–0.82) and 0.43 (HDI_{95%} = 0.31–0.55; PI_{95%} = 0.07–0.74). The production effect

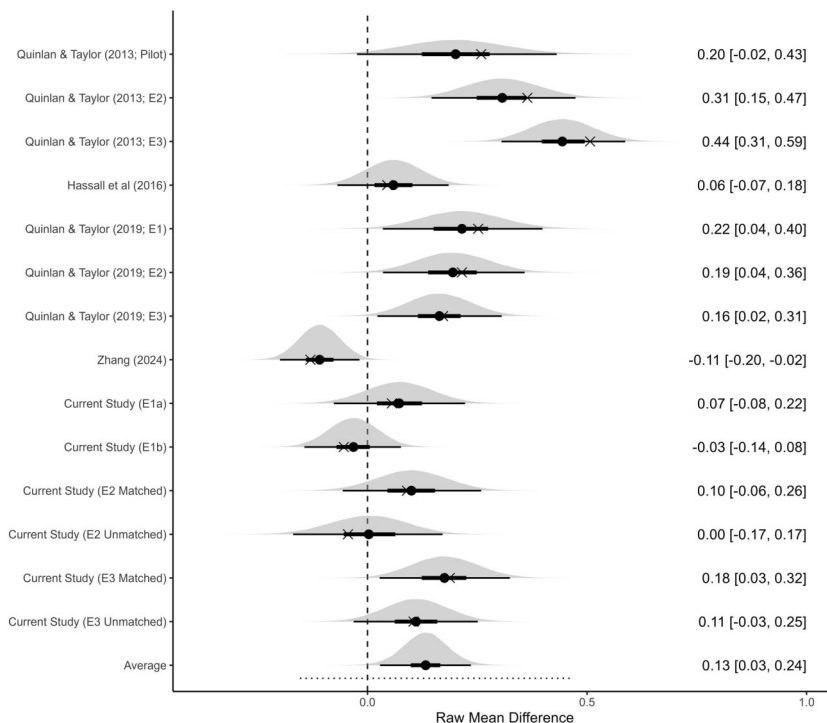


Figure 5. Forest plot depicting raw mean differences in d' (sing–aloud) for a meta-analytic model of the singing superiority effect. Note. Polygons depict the posterior distribution for each estimate, and points show the median estimate; observed effects are represented by an “X.” Thick lines represent the 50% HDI, and thin lines represent the 95% HDI. The dotted line represents the 95% PIs.

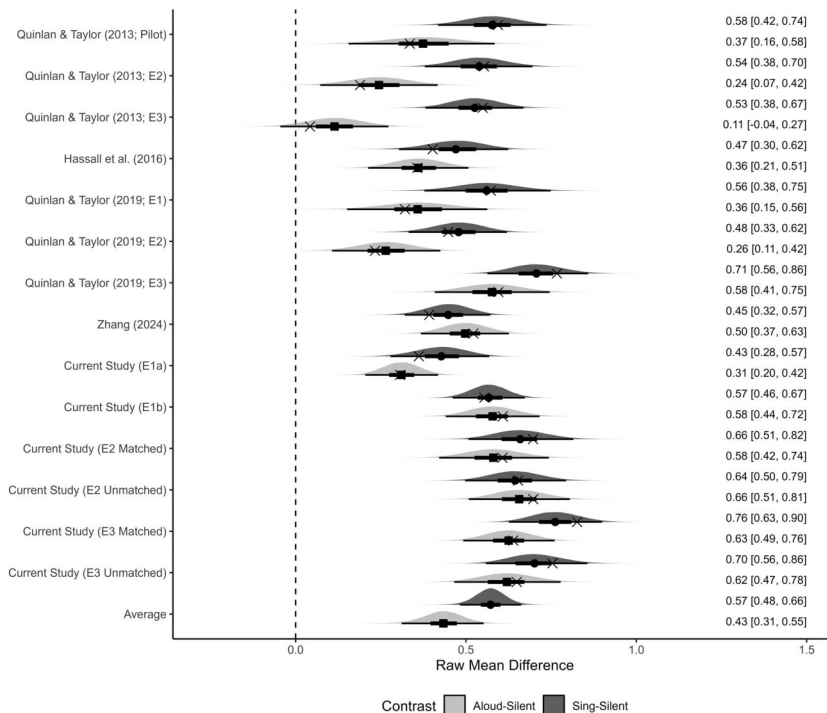


Figure 6. Forest plot depicting raw mean differences in d' (aloud–silent and sing–silent) for a meta-analytic model of the production effect. Note. Light-colored polygons and square points represent the difference in d' between the aloud and silent conditions, whereas dark-colored polygons and circular points represent contrasts between sing and silent. Polygons depict the posterior distribution for each contrast, and points show the median estimate; observed effects are represented by an “X.” Thick lines represent the 50% HDI, and thin lines represent the 95% HDI.

for singing was credibly larger than that for reading aloud, with the contrast between effects estimated at 0.14 ($HDI_{95\%} = 0.01\text{--}0.27$).

Subsequently, we fit an additional exploratory model using the approach outlined above to test for effects of color matching. Regardless of whether color matching was present, production effects for both singing and reading aloud were credible. However, the production effect for singing was credibly larger only for matched experiments, with the contrast between effects for this group estimated at 0.24 ($HDI_{95\%} = 0.08\text{--}0.39$). Conversely, the production effects for each condition were similar when color matching was not present (difference = 0.02, $HDI_{95\%} = -0.15$ to 0.20). A numerical trend favored a smaller production effect for reading aloud in color-matched experiments, with the difference between aloud conditions estimated at 0.15 ($HDI_{95\%} = -0.09$ to 0.37). Accordingly, both modeling approaches favor similar inferences. Although small, the aggregate SSE is credible; however, moderator analyses suggest that this effect is driven by larger SSEs in studies using color matching, whereas no credible effect appears to be present in studies that did not adopt this procedure.

General Discussion

The present study evaluated evidence favoring singing as a mnemonically superior production modality compared

to reading aloud. Across several conceptual and direct replications of previous work on the *singing superiority effect* (SSE) – as well as a meta-analysis of all published studies – we found the SSE to be smaller than previously thought and possibly dependent on color matching. In Experiments 1a and 1b, we explored the within-subject SSE using a standard three-condition design without matching test-phase font color to study phase condition. Experiments 2 and 3 instead provided a near replication of Quinlan and Taylor’s (2013) methods, including a group for whom test items were presented in the corresponding font color from study. Experiment 4 explored the SSE using a between-subject design. We observed robust production effects for both singing and reading aloud across all experiments but an SSE emerged only for the color-matched group in Experiment 3. A meta-analysis of all known studies demonstrated evidence of a small SSE, but moderator analyses likewise revealed this effect to be present only for color-matched conditions.

Based on these findings, support for the sensorimotor scaling hypothesis would appear to be limited. At the very least, the SSE appears to be much smaller than originally thought. For example, the initial observation made by Quinlan and Taylor (2013, Experiment 2) was a mean difference in sensitivity of ~ 0.36 , whereas the sole SSE we observed was estimated at only 0.23 and our meta-analytic estimate was 0.16 overall and 0.23 in studies using color matching. Critically, the large effect sizes

reported in Quinlan and Taylor (2013) were derived from small samples ranging from 15 to 22 participants. Later investigations by Quinlan and Taylor (2019) and Hassall et al. (2016) using larger samples (e.g., $N = 27\text{--}43$) reported smaller effect sizes better aligned with the differences observed in the present investigation (e.g., $MD = \sim 0.17$). Because smaller studies only have adequate statistical power to detect large effects, estimates derived from such samples are susceptible to overestimation (e.g., Sterne et al., 2000). Given that our meta-analytic model suggested that the aggregate benefit was driven largely by small studies that observed large effects, it appears likely that large effects previously reported reflect inflated estimates. Furthermore, there is no *a priori* theoretical basis to suggest that the SSE should be as large as previously reported. Because typical production benefits deriving from reading aloud already entail large benefits to sensitivity relative to silent reading (e.g., $MD = \sim 0.78$; Forrin et al., 2016), it seems unlikely that a more elaborate form of vocalization should nearly double the size of the effect.

In addition to being small in magnitude, the fact that the SSE emerges only for studies using color matching at test raises questions pertaining to the theoretical mechanisms at play, as well as the SSE's generalizability to other designs. One explanation might be that presenting items in their study colors permits different strategies to be employed across conditions. On the one hand, knowing a test item was studied silently might discourage reactivation or utilization of sensorimotor information, as it would not be expected. Separately, knowing a test item was sung or read aloud might encourage reactivation of very specific sensorimotor information. With respect to why an alternative type of distinctiveness heuristic might preferentially lead to a larger production effect for singing, it is possible that tonal or rhythmic information is useful in guiding retrieval, but that participants simply do not check for these features in typical paradigms. In this sense, information about stimulus dimensions derived from color matching might help focus the search for distinctive information on modality-specific features. For a benefit to emerge, the production trace must be utilized to guide retrieval; if participants typically neglect additional features specific to singing, no SSE would be expected.

An analogous alternative is that color matching could help reinstate context at test. Wakeham-Lewis et al. (2022) suggested that in production paradigms, participants might consciously reinstate the study phase production condition at to aid item discrimination (e.g., by thinking about saying the item aloud). The most natural

approach to doing so would be to imagine reading the item aloud in a normal speaking voice; however, unless prompted to do so, it is unlikely participants would imagine singing the item. According to this *sensorimotor reinstatement hypothesis*, recreating the productive act in one's mind would benefit singing (or other elaborate modes of speaking; Wakeham-Lewis et al., 2022). Providing cues about how an item would have been produced might guide participants to reinstate production in a manner attuned to study phase conditions. Much like our discussion above, such an explanation would suggest singing *does* encode additional information that drives superior memory relative to reading aloud, but that this information is useful only when heuristics atypical to production paradigms are applied to retrieve the information. Although distinctiveness- and context-based accounts provide plausible (albeit speculative) explanations for the interaction between the SSE and color matching at test, they do not necessarily provide a theoretical basis for why the effect does not reliably emerge even in paradigms that utilize this procedure: If singing encodes additional distinctive features relative to reading aloud, the effect should be robust across methodological variations. While it could be the case that the SSE is simply too small to detect reliably across experiments, our findings argue against this notion: The majority of our experiments failed to detect a credible effect despite using samples that were much larger than those used in previous efforts.

Perhaps the simplest explanation for our difficulty in replicating the SSE is that singing does not append additional distinctive features to the production trace relative to reading aloud. Quinlan and Taylor (2013, 2019; Hassall et al., 2016) argued that production via singing benefits from features related to pitch or tone. This is generally congruent with earlier literature, which has suggested that mnemonic benefits related to song derive because participants leverage melodic or rhythmic information in a process analogous to a distinctiveness heuristic (e.g., Wallace, 1994; but see Rainey & Larsen, 2002). However, the features thought to afford a relative benefit are not necessarily specific to singing: Human speech intrinsically incorporates varying degrees of rhythm, melody (e.g., Xu, 2005), pitch, and timbre (e.g., Dolson, 1994). If one accepts that all these features should also be present for items read aloud, the sensorimotor scaling hypothesis (e.g., Forrin et al., 2012) would not predict an SSE.

However, this account might be theoretically *rescued* if it allows for the possibility that variation in distinctive features can be qualitative rather than solely

quantitative. Rather than simply appending features that reflect additional distinct processes to the production trace, singing might allow for a greater degree of variation in item representations across features related to articulation and audition that are already present for speaking. Consistent with this notion, Caplan and Guitard (2024) propose a novel computational model of the production effect, which assumes both produced and unproduced items to be represented across phonological, orthographic, and semantic dimensions. However, produced items are thought to encode more phonological features, permitting additional variation in representations. Although both singing and reading aloud incorporate processing along tonal and rhythmic dimensions, these processes are utilized differently across modalities; for example, the use of pitch in singing is more precise and organized relative to speech (e.g., Zatorre & Baum, 2012). Accordingly, it is possible that singing could encode different features within the phonological subspace, leading to differences in item representations despite common sensory processing. Such a model might also accommodate an interaction between singing and color matching at test: Regardless of modality, produced items share common phonological features that participants may not normally distinguish between even if features related to singing possess additional discriminative value. However, participants might capitalize on the diagnostic value of this variation when prompted by cues at test to search for modality-specific information.

The present study is not the first to observe a non-significant SSE. Hassall et al. (2016) reported a pattern of results consistent with our initial experiments (i.e., sing = aloud > silent) despite using matched foils at test. Those authors explained their failure to replicate the effect with reference to methodological differences, suggesting that the effect did not emerge either because of a delay in production necessitated by their paradigm or because participants failed to tonally differentiate singing and speaking at study. However, neither of these explanations can satisfactorily account for our own failures to observe an SSE. Our experiments used standard production paradigms that did not separate productive cues and acts, indicating that any failures to replicate the effect could not be attributed to temporal separation. With respect to a *lazy singing* hypothesis, we

ensured that our participants in Experiments 1b onward were supervised throughout the study phase and prompted participants to sing more effortfully if their singing faltered. Furthermore, participants in Experiments 2–4 were assigned a rating by the experimenter, reflecting how effortfully each participant was tonally distinguishing their singing from regular speech. Although subjective, these ratings were typically high, suggesting that our participants did not simply fail to adequately sing the items at study.² Because previous efforts did not go as far as to implement these safeguards, it seems unlikely that our findings could be attributed to lack of participant effort.

However, it is also possible that simply singing effortfully is not sufficient for an SSE to emerge; rather, it may be the case that singing must also be sufficiently complex or varied. To this point, singing in real-world scenarios would not typically entail sequentially producing a list of unrelated words, as was the case in our paradigms. In this unusual situation, participants might be inclined to sing monotonically and with limited variation across melodic or rhythmic dimensions; producing items in this manner might amount to little more than emphasized speech, affording limited advantage. To address this possibility, ongoing behavioral and computational investigations in our own laboratory have been designed to evaluate the SSE in conjunction with manipulations designed to increase the variability of singing. For example, having participants produce short sentences might implicitly encourage tonal variation, as singing phrases is more natural than singing words. Similarly, providing explicit instructions to sing stimuli in accordance with melodies of varying complexity could shed light on the degree of tonal variation required to produce an SSE. While it is possible that an SSE could emerge under these conditions, the goal of the present work was to evaluate the SSE as it exists within the present literature: Nearly, all investigations prior to our own have reported a reliable advantage for singing, despite using paradigms and instructions that are nearly identical to those employed herein (e.g., Quinlan & Taylor, 2013, 2019). As such, there is little reason to suspect that participants in our own samples would have behaved differently.

With further respect to individual differences in the mnemonic utility of singing, it could also be that

² Participants were assigned a rating on a 10-point scale, with higher ratings indicating greater effort. For Experiment 2, the mean rating was 8.41 ($SD = 1.65$) in the matched group and 7.72 ($SD = 1.87$) in the unmatched group. For Experiment 3, the mean rating was 7.22 ($SD = 2.30$) in the matched group and 6.92 ($SD = 2.40$) in the unmatched group. Finally, the mean rating for participants in the sing condition in Experiment 4 was 7.44 ($SD = 2.26$). For each experiment, we also fit an exploratory model similar to those detailed in-text, albeit including singing ratings as a covariate. In all cases, no credible interactions between ratings of singing quality and sensitivity or response bias in any condition emerged.

embarrassment due to singing in front of an experimenter or lack of experience might have obfuscated an SSE for some participants. However, both past research and our own findings argue against this possibility. Previously, Quinlan and Taylor (2019; Experiment 1) targeted a sample of practiced singers, each of whom reported a minimum of one year of experience singing in front of an audience (e.g., as members of a choir). This experiment observed an SSE that was similar in size to other differences reported in that investigation ($MD = \sim 0.22$) and much smaller than earlier observations in inexperienced samples (e.g., $MD = \sim 0.43$; Quinlan & Taylor, 2013; Experiment 3). Furthermore, we explored the impact that singing in front of an experimenter might have in our own investigations by fitting an exploratory model comparing Experiments 1a and 1b, the former of which was unsupervised. Here, however, no credible differences in sensitivity for singing or the size of the SSE emerged.³ With this evidence in mind, it does not appear that effort, differences in instruction, experience, or embarrassment can account for the failure of the SSE to emerge in our study. While Hassall et al. (2016; also see Quinlan & Taylor, 2019) posited that their observation of a null effect was an atypical exception to a reliable advantage for singing, our findings instead suggest that this advantage is itself atypical and can emerge only when certain conditions are met.

Given the boundary conditions that the present study imposes on the SSE, our findings argue against claims that such an effect provides strong support for the sensorimotor scaling hypothesis (e.g., Quinlan & Taylor, 2013, 2019). Whereas such an account would predict the SSE to be driven by improved memory for singing, our findings cannot rule out the possibility that the effect might be driven by a decrement to performance for aloud items. When color matching at test was used, numerical trends in Experiment 2 and our meta-analysis favored lower sensitivity in the aloud condition while sensitivity in the sing condition remained comparable across groups. Furthermore, we observed no credible difference in sensitivity between conditions in Experiment 4, for which our design did not permit the emergence of within-participant *costs* to performance. Were the SSE facilitated by the addition of distinctive features to the production trace, the effect should not impair performance

for aloud items and should emerge irrespective of whether production is manipulated within- or between-subject.⁴ Rather, the pattern of results we observed suggests that participants might preferentially attend to or rehearse items sung at study, leading these items to be better represented in memory at the cost of poorer representations for aloud items.

Why participants would focus preferentially on singing is not clear, but this could occur due to perceived emphasis on singing during instruction or the inherent peculiarity of the modality relative to reading aloud (but see Quinlan & Taylor, 2019). Exploratory analyses of our own data and data provided by others provide some support for this hypothesis: Participants who exhibited an SSE generally exhibited performance for aloud items that was below average, while performance for singing for these participants was near the overall mean. If the SSE were to arise on the basis of such a mechanism, however, it would be attributable to preferential attentional allocation or rehearsal rather than sensorimotor scaling (for evidence that viewing certain production conditions as especially important may improve performance at the cost of other production conditions, see Ozubko et al., 2020).

In sum, the present investigation poses a challenge to the SSE as described in earlier literature (e.g., Quinlan & Taylor, 2013, 2019). Across most of our analyses, we observed a production effect for singing that was generally similar in magnitude to that for reading aloud (also see Hassall et al., 2016). When the SSE did emerge, it was much smaller than previous estimates and was confined to the color-matched group. Contrary to sensorimotor scaling explanations of the effect, it appears that the relative superiority of singing arises on the basis of idiosyncratic methodological factors. Even if these factors can be leveraged via an atypical distinctiveness heuristic or some alternative mechanism, it does not seem that singing affords any additional discriminative utility to the production trace that is immediately accessible in typical paradigms. Given that the SSE does not appear to arise solely on the basis of appending additional distinctive features to the production trace, our findings argue that the effect should not be construed as strong evidence for the sensorimotor scaling hypothesis.

³ A similar pattern of results was observed by Wakeham-Lewis et al. (2022), who had participants produce words using character voices. Much like singing, producing items in this manner in front of an experimenter could be considered embarrassing by some participants. However, Wakeham-Lewis et al. (2022) found no credible interaction between experimenter presence and the production effect for this modality; in fact, numerical trends suggested that the benefit tended to be larger when participants were supervised.

⁴ Although Quinlan and Taylor (2019) argued that between-subject production effects do not arise on the basis of encoding distinctiveness, participants report utilizing strategies resembling distinctiveness- or context-based heuristics in production paradigms, regardless of design (Fawcett & Ozubko, 2016).

Electronic Supplementary Material

The electronic supplementary materials are available with the online version of the article at <https://doi.org/10.1027/1618-3169/a000614>

ESM 1. Methodological details and supplementary analyses.

References

- Baayen, R. H., Tweedie, F. J., & Schreuder, R. (2002). The subjects as a simple random effect fallacy: Subject variability and morphological family effects in the mental lexicon. *Brain and Language, 81*(1-3), 55–65. <https://doi.org/10.1006/brln.2001.2506>
- Bodner, G. E., Jamieson, R. K., Cormack, D. T., McDonald, D. L., & Bernstein, D. M. (2016). The production effect in recognition memory: Weakening strength can strengthen distinctiveness. *Canadian Journal of Experimental Psychology, 70*(2), 93–98. <https://doi.org/10.1037/cep0000082>
- Bodner, G. E., Taikh, A., & Fawcett, J. M. (2014). Assessing the costs and benefits of production in recognition. *Psychonomic Bulletin & Review, 21*(1), 149–154. <https://doi.org/10.3758/s13423-013-0485-1>
- Bürkner, P. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software, 80*(1), 1–28. <https://doi.org/10.18637/jss.v080.i01>
- Caplan, J. B., & Guitard, D. (2024). A feature-space theory of the production effect in recognition. *Experimental Psychology, 71*(1), 64–82. <https://doi.org/10.1027/1618-3169/a000611>
- Conway, M. A., & Gathercole, S. E. (1987). Modality and long-term memory. *Journal of Memory and Language, 26*(3), 341–361. [https://doi.org/10.1016/0749-596X\(87\)90118-5](https://doi.org/10.1016/0749-596X(87)90118-5)
- Craik, F. I., & Lockhart, R. S. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning & Verbal Behavior, 11*(6), 671–684. [https://doi.org/10.1016/S0022-5371\(72\)80001-X](https://doi.org/10.1016/S0022-5371(72)80001-X)
- Cyr, V., Poirier, M., Yearsley, J. M., Guitard, D., Harrigan, I., & Saint-Aubin, J. (2022). The production effect over the long term: Modeling distinctiveness using serial positions. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 48*(12), 1797–1820. <https://doi.org/10.1037/xlm0001093>
- 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*(1), 155–161. <https://doi.org/10.3758/BF03196152>
- Dolson, M. (1994). The pitch of speech as a function of linguistic community. *Music Perception, 11*(3), 321–331. <https://doi.org/10.2307/40285626>
- Fawcett, J. M. (2013). The production effect benefits performance in between-subject designs: A meta-analysis. *Acta Psychologica, 142*(1), 1–5. <https://doi.org/10.1016/j.actpsy.2012.10.001>
- Fawcett, J. M., Baldwin, M. M., Whitridge, J. W., Swab, M., Malayang, K., Hiscock, B., Drakes, D. H., & Willoughby, H. V. (2023). Production improves recognition and reduces intrusions in between-subject designs: An updated meta-analysis. *Canadian Journal of Experimental Psychology, 77*(1), 35–44. <https://doi.org/10.1037/cep0000302>
- Fawcett, J. M., Lawrence, M. A., & Taylor, T. L. (2016). The representational consequences of intentional forgetting: Impairments to both the probability and fidelity of long-term memory. *Journal of Experimental Psychology: General, 145*(1), 56–81. <https://doi.org/10.1037/xge0000128>
- 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*(2), 99–115. <https://doi.org/10.1037/cep0000089>
- Fawcett, J. M., Quinlan, C. K., & Taylor, T. L. (2012). Interplay of the production and picture superiority effects: A signal detection analysis. *Memory (Hove), 20*(7), 655–666. <https://doi.org/10.1080/09658211.2012.693510>
- Fernandes, M. A., Wammes, J. D., & Meade, M. E. (2018). The surprisingly powerful influence of drawing on memory. *Current Directions in Psychological Science, 27*(5), 302–308. <https://doi.org/10.1177/0963721418755385>
- Forrin, N. D., Groot, B., & MacLeod, C. M. (2016). The d-prime directive: Assessing costs and benefits in recognition by dissociating mixed-list false alarm rates. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 42*(7), 1090–1111. <https://doi.org/10.1037/xlm0000214>
- Forrin, N. D., & MacLeod, C. M. (2018). This time it's personal: The memory benefit of hearing oneself. *Memory, 26*(4), 574–579. <https://doi.org/10.1080/09658211.2017.1383434>
- Forrin, N. D., MacLeod, C. M., & Ozubko, J. D. (2012). Widening the boundaries of the production effect. *Memory & Cognition, 40*(1), 1046–1055. <https://doi.org/10.3758/s13421-012-0210-8>
- Gathercole, S. E., & Conway, M. A. (1988). Exploring long-term modality effects: Vocalization leads to best retention. *Memory & Cognition, 16*(2), 110–119. <https://doi.org/10.3758/BF03213478>
- Gionet, S., Guitard, D., & Saint-Aubin, J. (2024). The production effect interacts with serial positions in recall tasks, but not in item recognition. [Manuscript submitted for publication]. *Experimental Psychology*.
- Hassall, C. D., Quinlan, C. K., Turk, D. J., Taylor, T. L., & Krigolson, O. E. (2016). A preliminary investigation into the neural basis of the production effect. *Canadian Journal of Experimental Psychology, 70*(2), 139–146. <https://doi.org/10.1037/cep0000093>
- Hopkins, R. H., & Edwards, R. E. (1972). Pronunciation effects in recognition memory. *Journal of Verbal Learning and Verbal Behavior, 11*(4), 534–537. [https://doi.org/10.1016/S0022-5371\(72\)80036-7](https://doi.org/10.1016/S0022-5371(72)80036-7)
- Huff, M. J. (2019, November 11). *The effects of singing on enhancing the Production Effect* [Pre-Registration]. <https://doi.org/10.17605/OSF.IO/Z6JUE>
- Icht, M., Bergerzon-Biton, O., & Mama, Y. (2019). The production effect in adults with dysarthria: Improving long-term verbal memory by vocal production. *Neuropsychological Rehabilitation, 29*(1), 131–143. <https://doi.org/10.1080/09602011.2016.1272466>
- IntHout, J., Ioannidis, J. P., Rovers, M. M., & Goeman, J. J. (2016). Plea for routinely presenting prediction intervals in meta-analysis. *BMJ Open, 6*(7), Article e010247. <https://doi.org/10.1136/bmjopen-2015-010247>
- Isarida, T., & Isarida, T. K. (2007). Environmental context effects of background color in free recall. *Memory & Cognition, 35*(7), 1620–1629. <https://doi.org/10.3758/BF03193496>
- Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language, 59*(4), 434–446. <https://doi.org/10.1016/j.jml.2007.11.007>
- 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*(2), 154–164. <https://doi.org/10.1037/cep0000081>
- jed709 (2024). *jed709/ExPsy-Production-Singing: Production-Singing-Open-Data (v1.1)* [Data]. Zenodo. <https://zenodo.org/records/11205433>

- Kelly, M. O., Ensor, T. M., Lu, X., MacLeod, C. M., & Risko, E. F. (2022). Reducing retrieval time modulates the production effect: Empirical evidence and computational accounts. *Journal of Memory and Language*, 123(1), Article 104299. <https://doi.org/10.1016/j.jml.2021.104299>
- Kruschke, J. K. (2010). Bayesian data analysis. *Wiley Interdisciplinary Reviews: Cognitive Science*, 1(5), 658–676. <https://doi.org/10.1002/wcs.72>
- Lin, O. Y. H., & MacLeod, C. M. (2012). Aging and the production effect: A test of the distinctiveness account. *Canadian Journal of Experimental Psychology*, 66(3), 212–216. <https://doi.org/10.1037/a0028309>
- 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(3), 671–685. <https://doi.org/10.1037/a0018785>
- Mama, Y., & Icht, M. (2016). Auditioning the distinctiveness account: Expanding the production effect to the auditory modality reveals the superiority of writing over vocalising. *Memory*, 24(1), 98–113. <https://doi.org/10.1080/09658211.2014.986135>
- Masson, M. E. (2011). A tutorial on a practical Bayesian alternative to null-hypothesis significance testing. *Behavior Research Methods*, 43(1), 679–690. <https://doi.org/10.3758/s13428-010-0049-5>
- McCurdy, M. P., Viechtbauer, W., Sklenar, A. M., Frankenstein, A. N., & Leshikar, E. D. (2020). Theories of the generation effect and the impact of generation constraint: A meta-analytic review. *Psychonomic Bulletin & Review*, 27(6), 1139–1165. <https://doi.org/10.3758/s13423-020-01762-3>
- Morey, R. D., Hoekstra, R., Rouder, J. N., Lee, M. D., & Wagenmakers, E. J. (2016). The fallacy of placing confidence in confidence intervals. *Psychonomic Bulletin & Review*, 23(1), 103–123. <https://doi.org/10.3758/s13423-015-0947-8>
- Ozubko, J. D., Bamburoski, L. D., Carlin, K., & Fawcett, J. M. (2020). Distinctive encodings and the production effect: Failure to retrieve distinctive encodings decreases recollection of silent items. *Memory (Hove)*, 28(2), 237–260. <https://doi.org/10.1080/09658211.2019.1711128>
- Ozubko, J. D., Gopie, N., & MacLeod, C. M. (2012). Production benefits both recollection and familiarity. *Memory & Cognition*, 40(3), 326–338. <https://doi.org/10.3758/s13421-011-0165-1>
- 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(6), 1543–1547. <https://doi.org/10.1037/a0020604>
- Peirce, J. W., Gray, J. R., Simpson, S., MacAskill, M. R., Höchenberger, R., Sogo, H., Kastman, E., & Lindeløv, J. (2019). PsychoPy2: Experiments in behavior made easy. *Behavior Research Methods*, 51(1), 195–203. <https://doi.org/10.3758/s13428-018-01193-y>
- Quinlan, C. K., & Taylor, T. L. (2013). Enhancing the production effect in memory. *Memory (Hove)*, 21(8), 904–915. <https://doi.org/10.1080/09658211.2013.766754>
- Quinlan, C. K., & Taylor, T. L. (2019). Mechanisms underlying the production effect for singing. *Canadian Journal of Experimental Psychology*, 73(4), 254–264. <https://doi.org/10.1037/cep0000179>
- R Core Team (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. <https://www.R-project.org/>
- Rainey, D. W., & Larsen, J. D. (2002). The effect of familiar melodies on initial learning and long-term memory for unconnected text. *Music Perception*, 20(2), 173–186. <https://doi.org/10.1525/mp.2002.20.2.173>
- Richler, J. J., Palmeri, T. J., & Gauthier, I. (2013). How does using object names influence visual recognition memory? *Journal of Memory and Language*, 68(1), 10–25. <https://doi.org/10.1016/j.jml.2012.09.001>
- Saint-Aubin, J., Yearsley, J. M., Poirier, M., Cyr, V., & Guitard, D. (2021). A model of the production effect over the short-term: The cost of relative distinctiveness. *Journal of Memory and Language*, 118(1), Article 104219. <https://doi.org/10.1016/j.jml.2021.104219>
- Slamecka, N. J., & Graf, P. (1978). The generation effect: Delineation of a phenomenon. *Journal of Experimental Psychology: Human Learning and Memory*, 4(6), 592–604. <https://doi.org/10.1037/0278-7393.4.6.592>
- Stanislaw, H., & Todorov, N. (1999). Calculation of signal detection theory measures. *Behavior Research Methods, Instruments, & Computers*, 31(1), 137–149. <https://doi.org/10.3758/BF03207704>
- Sterne, J. A., Gavaghan, D., & Egger, M. (2000). Publication and related bias in meta-analysis: power of statistical tests and prevalence in the literature. *Journal of Clinical Epidemiology*, 53(11), 1119–1129. [https://doi.org/10.1016/S0895-4356\(00\)00242-0](https://doi.org/10.1016/S0895-4356(00)00242-0)
- Taitelbaum-Swead, R. T., Mama, Y., & Icht, M. (2018). The effect of presentation mode and production type on word memory for hearing impaired signers. *Journal of the American Academy of Audiology*, 29(10), 875–884. <https://doi.org/10.3766/jaaa.17030>
- Tulving, E., & Thomson, D. M. (1973). Encoding specificity and retrieval processes in episodic memory. *Psychological Review*, 80(5), 352–373. <https://doi.org/10.1037/h0020071>
- Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, 36(3), 1–48. <https://doi.org/10.18637/jss.v036.i03>
- Wakeham-Lewis, R. M., Ozubko, J., & Fawcett, J. M. (2022). Characterizing production: The production effect is eliminated for unusual voices unless they are frequent at study. *Memory*, 30(10), 1319–1333. <https://doi.org/10.1080/09658211.2022.2115075>
- Wallace, W. T. (1994). Memory for music: Effect of melody on recall of text. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 20(6), 1471–1485. <https://doi.org/10.1037/0278-7393.20.6.1471>
- Wammes, J. D., Meade, M. E., & Fernandes, M. A. (2016). The drawing effect: Evidence for reliable and robust memory benefits in free recall. *Quarterly Journal of Experimental Psychology*, 69(9), 1752–1776. <https://doi.org/10.1080/17470218.2015.1094494>
- Williams, D. R., Rast, P., & Bürkner, P. (2018). *Bayesian meta-analysis with weakly informative prior distributions*. PsyArXiv. <https://doi.org/10.31234/osf.io/7tbrm>
- Wilson Van Voorhis, C. R., & Morgan, B. L. (2007). Understanding power and rules of thumb for determining sample sizes. *Tutorials in Quantitative Methods for Psychology*, 3(2), 43–50. <https://doi.org/10.20982/tqmp.03.2.p043>
- Xu, Y. (2005). Speech melody as articulatorily implemented communicative functions. *Speech Communication*, 46(3), 220–251. <https://doi.org/10.1016/j.specom.2005.02.014>
- Yonelinas, A. P. (2002). The nature of recollection and familiarity: A review of 30 years of research. *Journal of Memory and Language*, 46(3), 441–517. <https://doi.org/10.1006/jmla.2002.2864>
- Yonelinas, A. P., & Jacoby, L. L. (1995). The relation between remembering and knowing as bases for recognition: Effects of size congruency. *Journal of Memory and Language*, 34(5), 622–643. <https://doi.org/10.1006/jmla.1995.1028>
- Zatorre, R. J., & Baum, S. R. (2012). Musical melody and speech intonation: Singing a different tune. *PLoS Biology*, 10(7), Article e1001372. <https://doi.org/10.1371/journal.pbio.1001372>
- Zhang, B. (2024). *Comparing memory levels between reading aloud and singing*. Unpublished manuscript.

History

Received December 13, 2023

Revision received May 16, 2024

Accepted May 30, 2024

Published online July 30, 2024

Acknowledgments

An earlier version of this paper was presented at the 32nd annual meeting of the Canadian Society for Brain, Behaviour and Cognitive Science (Halifax, NS, Canada, July 2022).

Conflict of Interest

The authors declare no competing financial interests.

Publication Ethics

Informed consent was obtained from all participants included in the study.

Authorship

All authors approved the final version of the article.


Open Science

To the best of my ability and knowledge, I have provided all original materials and clear references to all other materials via a stable online repository. Several experiments reported in this manuscript were pre-registered. The pre-registration and analysis plan are available at <https://osf.io/z6jue> (Huff, 2019). Raw data and analysis


scripts are available at <https://doi.org/10.5281/zenodo.11205433> (jed709, 2024).

ORCID

Jedidiah W. Whitridge

 <https://orcid.org/0000-0003-1237-4977>


Mark J. Huff

 <https://orcid.org/0000-0002-0155-7877>

Chelsea D. Lahey

 <https://orcid.org/0000-0003-0585-8615>

Jonathan M. Fawcett

 <https://orcid.org/0000-0002-4248-5371>

Jedidiah Whitridge

Department of Psychology
Memorial University of Newfoundland
230 Elizabeth Avenue
St. John's
NL A1B3X1
Canada
jww828@mun.ca