**PSYC 6003: Final Paper**

Jedidiah W. Whitridge

Department of Psychology, Memorial University of Newfoundland

**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., tone). However, the singing superiority effect has not always replicated. Our investigation demonstrated a production effect for items read aloud but failed to observe a singing superiority effect, although numerical trends suggested that the relative superiority of singing might occur only when items are tested in the same colour in which they were studied (with foils randomized across colours). 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 colour in which they were studied. This outcome is inconsistent with common distinctiveness-based theoretical accounts.

*Keywords*: production, memory, singing, distinctiveness

**Final Assignment**

We rely on our memories in nearly all facets of life, from basic survival behaviors to higher learning and spirituality. It is no surprise, then, that a great deal of cognitive research has focused on strategies such as generation (e.g., Slamecka & Graf, 1978; McCurdy et al., 2020) or levels of processing (e.g., Craik & Lockhart, 1972) capable of improving memory for important information. 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 which has been termed the *production effect* (MacLeod et al., 2010). This benefit 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 & Mama, 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; Kappel et al., 1973). Although these processes remain a subject of debate (e.g., Fawcett, 2013; Fawcett et al., 2022), 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, with researchers demonstrating that 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 this 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; see also, 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, 1987). Whereas reading aloud incorporates visual, motoric, and auditory features, writing and mouthing exclude the auditory component. 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 favouring the sensorimotor scaling hypothesis is the finding that the incorporation of tonal and rhythmic information – via singing – produces an especially large production effect (e.g., Quinlan & Taylor, 2013, 2019). This *singing superiority effect* (SSE) was first demonstrated by Quinlan and Taylor (2013), who observed a greater benefit for singing compared to reading aloud across several experiments. Quinlan and Taylor (2019) subsequently replicated and extended this finding, 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 strong evidence of both the sensorimotor scaling hypothesis and the distinctiveness account more broadly (e.g., Forrin & MacLeod, 2018; Mama & Icht, 2016).

However, several efforts to replicate the SSE have failed (e.g., Hassall et al., 2016; Ozubko et al., 2020). Most recently, experiments from our own laboratory observed production effects for singing and reading aloud that were of similar magnitude (i.e., sing = aloud > silent). Before drawing strong conclusions about the reliability of the SSE, we considered the possibility that the effect might be driven by aspects of study design. Specifically, our earlier experiments deviated from the methods used by Quinlan and Taylor (2013), with one potentially important difference being those authors’ use of colour-matching at test (i.e., presenting test items in the same colour they were studied in with foils randomly intermixed between assignments). This procedure was first used by Fawcett et al. (2012) to permit separate false alarm rates and thereby the calculation of response bias and sensitivity 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 ought to 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).

However, we speculated that this decision might have obfuscated the SSE. Some variants of the distinctiveness account contend that participants leverage distinctive information about having produced information consciously in the form of a *distinctiveness heuristic*: At test, participants evaluate whether they believe they might have recently produced each item and use this record to guide discrimination (i.e., “I remember saying it aloud, so I must have studied it”; Dodson & Schacter, 2001; see also, MacLeod et al., 2010). By orienting participants to stimulus dimensions via item colour, it is possible that participants might employ different heuristics for each condition. For example, recognizing that test items presented in red were sung at study might lead participants to focus on tonal information they might have otherwise ignored had they not known the study phase condition. To explore this possibility, we conducted an experiment that conceptually replicated the methods of Quinlan and Taylor (2013; Experiment 2) and incorporated colour-matched items at test as a between-subject manipulation. Additionally, we report a meta-analysis of all known studies of the SSE to address discrepancy amongst reported effects.

To foreshadow our results, we observed a robust production effect for singing and reading aloud across all dependent measures, replicating our earlier investigations. However, only a non-credible numerical trend favored an SSE for the colour-matched group, with these findings hinting at the possibility that informing participants of the study phase condition of a given test item may facilitate observation of 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 colour-matching procedure.

**Method**

**Participants**

Participants consisted of 90 students (*N* = 45 matched) from The University of Southern Mississippi who took part in the experiment in exchange for partial course credit.

**Stimuli and Apparatus**

Stimuli were randomly selected from a pool of 360 words retrieved from the MRC Psycholinguistic Database (Coltheart, 1981). All words were nouns between 5 and 10 letters long and had SUBTLEX frequency scores between 0.41 and 4008.39 (*M* = 83.27, *SD* = 253.79).

Each participant was assigned a random subset of 180 words. Half (90 items) were studied by participants and were randomized between the three study conditions (i.e., 30 words each to read silently, read aloud, and sing). Words at study were presented in colored font, with each respective study condition being assigned either red, yellow or blue; color assignments were counterbalanced. All words were displayed in their assigned color at study. The remainder of the words (90 items) appeared only as “new” foils at test. For the matched group, test items were presented in their corresponding study phase colours, with foil items split equally across colours. For the unmatched group, all test items were presented in white font. All stimuli were presented in 14-point Arial font against a black background. The experimental paradigm was coded in PsychoPy (version 2.3.2; Peirce et al., 2019) and presented via a 20-inch color monitor attached to a computer running Windows 10.

**Procedure**

The experiment consisted of a study phase and a test phase. Prior to the study phase, participants were informed that they would see words presented in one of three colors (red, yellow, or blue) and that the color indicated how the words should be studied. The experimenter provided a demonstration to each participant for how words should be produced, with particular emphasis on the singing condition. Participants 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. As indicated by their color assignment, one third of the items were to be sung aloud, one third were to be read aloud and the remaining third were to be read silently (30 items each). 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 was present throughout the study phase. The experimenter monitored study responses in order 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 180 words, 90 of which were “old” words seen in the study phase and 90 of which were “new” foil words. For the matched group, all test items were presented in their corresponding study phase colours, (e.g., a word presented in blue at study was presented in blue at test) with foil items split equally across colour assignments. For the unmatched group, all test words were presented in white font. There were no other differences between the matched and unmatched groups. Each test trial 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 judgement and a recollect/familiar/neither judgement, which were separated by a 500 ms blank screen.

Confidence judgements 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 judgement was analogous to commonly employed remember/know/no judgements (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**

Rather than analyzing raw or corrected hit rates, we opted to estimate sensitivity (*d’*) and response bias (*C*), parameters derived from *signal detection theory* as applied to recognition memory (Egan, 1958). In this context, signal detection models contend that participants assess the familiarity of test items to decide whether to make an “old” or “new” response. If familiarity with a given item test exceeds the participant’s decision criterion, the participant will make an “old” response. Both studied and unstudied items are expected to elicit some degree of familiarity, although the degree to which a given stimulus is familiar will vary across trials. Accordingly, distributions of familiarity arise for both studied items (i.e., a “signal” distribution) and unstudied items (i.e., a “noise” distribution), with the mean of the former assumed to be higher than the latter due to recent exposure to studied items. Our parameters of interest can be derived from this general model, with *C* reflecting the decision criterion at which an “old” response will be made and *d’* reflecting the distance between the signal and noise distributions. Put simply, *C* quantifies participants’ propensity to identify items as “old” regardless of whether the items were studied, whereas *d’* quantifies propensity to successfully discriminate between old and new items (for a detailed overview of signal detection theory in the context of recognition memory, see, e.g., Macmillan & Creelman, 2005; see also, Banks, 1970; Stanislaw & Todorov, 1999). Our decision to adopt a signal detection approach was motivated by evidence that analysis of recognition memory using only raw or corrected hit rates often underestimates participants’ capacity for discrimination (e.g., Rouder et al., 2007; Stanislaw & Todorov, 1999; see Macmillan & Creelman, 2005, for a detailed discussion). Furthermore, several studies have argued for the superiority of signal detection analysis over raw hits for interpretation of the production effect on the basis that response bias is more liberal for produced relative to unproduced items; thus, failing to account for this parameter may overestimate the size of the effect and bias inference (e.g., Fawcett et al., 2012, 2023).

With this framework in mind, we opted to use Bayesian probit regression to estimate *d’* and *C* in a multilevel context. We view this approach as advantageous for several reasons. First, our primary dependent measures were binary.[[1]](#footnote-1) While binary outcome data (e.g., “old” or “new” responses) is often aggregated into proportions, this procedure violates assumptions made by typical statistical approaches (e.g., ANOVA) and is liable to inflate Type I error rates (e.g., Baayen et al., 2002; Jaeger, 2008). Conventional procedures may also be used to calculate *d’* and *C* by collapsing trial data into proportions and applying transformations (e.g., Stanislaw & Todorov, 1999), but this approach ignores variability that arises at the level of the trial or the participant and may bias inference: Aggregation of data across items and participants assumes that sensitivity and criteria are uniform across these parameters (Rouder et al., 2007; see also, Gelman & Hill, 2006). Such an assumption is highly unlikely to be true, given that a number of item-level characteristics (e.g., word frequency; Broadbent, 1967; concreteness; Paivio et al., 1994) and participant-level characteristics (e.g., age; Grady et al., 1995) have long been known to impact memory. However, these issues can be mitigated by using generalized linear mixed models to estimate relevant parameters whilst simultaneously accounting for sources of variability (Rouder & Lu, 2005; Rouder et al., 2007; Wright et al., 2009). Finally, the present experiment aimed to evaluate evidence favouring the existence of an effect, which must therefore accept the possibility that no effect exists. While typical Frequentist approaches to hypothesis testing allow only for failure to reject the null hypothesis on the basis of a null effect, Bayesian approaches allow for the quantification of evidence favoring a null model (Masson, 2011).

For these reasons, the analyses reported herein utilized multilevel probit regression models implemented via the *brms* package (Bürkner, 2017) in *R* (R Core Team, 2020; see also, Fawcett & Ozubko, 2016; Fawcett et al., 2016). Our decision to use probit models deviates slightly from comparable approaches within the production literature that have utilized logistic regression to estimate analogous parameters (e.g., Fawcett & Ozubko, 2016; Fawcett et al., 2016; for a comparable Frequentist approach, see Zormpa et al., 2019). However, logistic and probit signal detection models produce near identical estimates that differ only insofar as they exist on different scales (DeCarlo, 1998). Because conventional calculations of *d’* and *C* utilize probit transformations (e.g., Stanislaw & Todorov, 1999), the estimates for these parameters produced by probit regression exist on the same scale as those produced by conventional procedures. Thus, we opted for probit regression over logistic regression to ensure that our estimates would be easily interpreted by readers familiar with signal detection theory.

Our models were parameterized in accordance with the general structure for generalized linear signal detection models outlined by DeCarlo (1998; see also, Rouder et al., 2007; Stanislaw & Todorov, 1999; Wright et al., 2009), albeit adapted for probit models using a Bayesian approach (for a tutorial, see Vuorre, 2017). For this parameterization, the probability of a correct “old” response (i.e., a *hit*), *H*,is given by

and the probability of an incorrect “old” response (i.e., a *false alarm*), *F*, is given by

where Φ denotes the normal cumulative distribution function (i.e., *z*-transformed probabilities). To estimate the overall probability, *p*, of an “old” response for a given item, *i*, let *old* denote a dummy-coded binary classification indicating whether an item was studied. The probability of an “old” response for the *i*th item is then given by

Thus, the probability of a hit is calculated when *old* = 1, and the probability of a false alarm is calculated when *old* = 0.[[2]](#footnote-2) All models were parameterized using the general structure given by this nonlinear equation, albeit also including fixed and random effects. This parameterization permitted *d’* and *C* to be estimated directly from the data, with fixed and random effects applied independently to each parameter. Using this approach, we removed the model intercept and computed slopes that produced estimates of *d’* and *C* for all possible combinations of our fixed effects. These parameters were estimated for each of our dependent measures (i.e., confidence, recollection, and familiarity) using separate models.

Each model included fixed effects for condition (sing, aloud, silent) and group (matched, unmatched) on both *d’* and *C*. With respect to our random effects structure, we assumed that the impact of our fixed effects would vary across participants and items; this structure was adopted because failure to account for item- and participant-level variability in effects – or accounting solely for baseline levels variability along these dimensions (i.e., by including only a random intercept) – can produce biased estimates (Rouder et al., 2007). Thus, we included random slopes to permit item-level variation in the impact of the fixed effects of group and condition. For participant-level effects, however, our random slopes only permitted variation in the effect of condition, given that group was manipulated between-subjects and participant-level effects corresponding to this parameter were thereby not justified by the design (see, e.g., Barr et al., 2013; Gelman & Hill, 2006). Although we removed the intercept from our models in all cases, we modeled correlations between random slopes reflecting our assumption of baseline item- and participant-level variability in *d’* and *C*.[[3]](#footnote-3) While not all model terms corresponding to our random effects are reported in-text, an overview of these estimates is provided for each model reported below.

For each model, we applied uninformative, mildly regularizing priors. Our priors were specified to reflect our belief that sensitivity for any given condition (i.e., sing, aloud, silent) should reasonably fall between -1 and 3 and that response bias should fall between -2 and 2; these priors were calibrated with respect to effects observed in other signal detection analyses of the production effect (e.g., Fawcett et al., 2012; Forrin et al., 2016; Quinlan & Taylor, 2013). For random effects, we applied priors calibrated to beliefs that the standard deviation for any given clustering variable across these parameters should fall between 0 and 2. Finally, where applicable, we also applied mildly regularizing *lkj* priors to correlations between random effects with a scale of 4 (Lewandowski et al., 2009; for further discussion, see McElreath, 2018).

All models for this experiment were fit using 8 independent chains of 15000 iterations each with a warm-up period of 7500 iterations. Model convergence was assessed using R-hat statistics, which were less than 1.00 in all cases, indicating that all models converged (Gelman & Hill, 2006; Kruschke, 2010). Further, inspection of the chains showed that effective sample size was greater than 7000 for all estimates reported in-text or in this supplement and greater than 10000 in most cases. Chains for some item-level correlation terms were less efficient, but effective sample size for these coefficients was greater than 3000 in all cases.

For each model, we report median posterior estimates for *d’* by condition and group and contrasts between conditions by group. The latter parameters were calculated directly from the posterior distributions of the estimates for each condition and reflect raw differences in each parameter. 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). This interval quantifies uncertainty around the posterior estimate and can be used to adjudicate whether estimates are credibly different from zero, analogous to statistical significance in Frequentist analysis. For example, if 95% of credible values are above zero, this can be interpreted as indicating 95% confidence that the estimate is positive. On the other hand, a 95% HDI that contains zero represents an estimate that is not credibly different from zero.

With respect to our analyses of response bias, our colour matching procedure allowed us to record separate false alarm rates that permitted direct, condition-specific estimates of *C* for the matched group. However, estimates of *C* for the unmatched group were calculated using arbitrarily separated false alarm rates. Because the calculation of *C* uses both hit rates and false alarm rates, condition-specific estimates of this parameter for the unmatched group capture differences in hit rates but not false alarm rates. As a result, estimates of response bias can be meaningfully interpreted only for the matched group; our discussion of this parameter thereby focuses solely on the matched group. With this limitation in mind, we also report median posterior estimates for *C* by condition and group and contrasts between conditions by group; the 95% HDI is reported alongside each estimate. Estimates for this parameter can be interpreted such that lower values reflect more liberal response bias (i.e., a higher propensity to respond with “old” irrespective of whether an item is old), whereas higher values indicate more conservative responses.

**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 binarized responses with condition (sing, aloud, silent) and group (matched, unmatched) as fixed effects.

*Sensitivity.* As depicted in Figure 1, 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 = 0.16 (HDI95% = -0.17 – 0.45), hinting at a potential interaction between singing and colour matching.

Surprisingly, our analyses of confidence ratings failed to detect the singing superiority effect reported in Quinlan and Taylor (2013, 2019). Instead, we observed a production effect for singing that was similar in magnitude to that for reading aloud, a pattern of results akin to those reported by Hassall et al. (2016). These findings appear to pose an initial challenge to the reliability of the singing superiority effect.

As outlined above, our design did not justify the inclusion of participant-level random effects for group. Thus, estimates corresponding to participant-level random effects reflect variability in both the matched and unmatched groups. With this in mind, participant-level random slopes corresponding to the effect of condition were informative in all cases (all estimates > 0.20), indicating variability in the impact that study modality had on sensitivity across individuals. Additionally, numerical trends indicated more participant-level variability in sensitivity in either production condition relative to the silent condition. Participant-level correlations between the aloud/sing and sing/silent conditions were moderate and positive (estimates > 0.45). The correlation between the aloud and silent conditions was smaller but still positive, although this trend failed to reach credibility (estimate = 0.33, HDI95% = -0.11 – 0.72). Nonetheless, the results generally suggest that participants with higher sensitivity in one condition also exhibited higher sensitivity in other conditions, indicative of baseline variability in this parameter across participants.

Item-level random slopes corresponding to the effects of condition and group were also informative in all cases (estimates > 0.20 for the matched group and > 0.28 for the unmatched group). Again, numerical trends favored higher variability in the impact of either production condition relative to the silent condition. Interestingly, trends also favored higher variability in the effect of all conditions on sensitivity for the matched group relative to the unmatched group. For item-level effects, correlation terms were less informative albeit positive in all cases, suggesting some degree of baseline variation in sensitivity across items. Overall, the trends observed support the random effects structure we chose for our models.

*Response Bias.* As shown in Figure 2, response bias was credibly more liberal for either production condition relative to the silent condition. This finding is generally congruent with earlier literature that has identified liberal shifts in response bias for produced relative to silent items (e.g., Fawcett et al., 2012; Quinlan & Taylor, 2013; Zormpa et al., 2019) and highlights the superiority of signal detection analysis for production studies. Interestingly, response bias was credibly more conservative for the sing relative to aloud condition (difference = 0.12, HDI95% = 0.01 – 0.23). This trend is novel, with earlier studies observing similar response bias for sing and aloud items (Quinlan and Taylor, 2013). Why this pattern has been hitherto unobserved is unclear, although the effect may simply have emerged due to the relatively greater statistical power of the present experiment.

Because participant-level effects could not be clustered by group and because of the limitations associated with calculating *C* for the unmatched group (as described above), all participant-level random effects on response bias reported hereafter should be interpreted with caution. With this caveat in mind, participant-level random slopes corresponding to the effect of condition were informative in all cases (all estimates > 0.48), indicating variability in the impact that study modality had on response bias across individuals. Participant-level correlations between all conditions were strong and positive (estimates > 0.76), reflecting baseline variability in response bias across participants.

Item-level random slopes corresponding to the effects of condition and group were also informative in all cases and similar across groups (estimates > 0.29 for the matched group and > 0.28 for the unmatched group). For response bias, item-level correlations were moderate, positive, and similar across groups (estimates > 0.48 for the matched group and > 0.55 for the unmatched group), indicating baseline variation in response bias across items. Overall, these trends are generally consistent with those observed in the random effects for sensitivity.

**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.

*Sensitivity.* As shown in Figure 1, we observed production effects for both singing and reading aloud that were credibly different from zero and of similar size to the production effects observed for confidence ratings. A credible SSE on recollection failed to emerge for either group, although a numerical trend once again favored a slightly larger advantage for singing in the matched group (difference = 0.18, HDI95% = -0.18 – 0.53). As expected, this model largely 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.

For our analysis of recollection, participant-level random slopes corresponding to the effect of condition were informative in all cases (all estimates > 0.30), indicating variability in the impact that study modality had on sensitivity across individuals. Like our analysis of confidence ratings, numerical trends suggested more participant-level variability in sensitivity for either production condition relative to the silent condition. Participant-level correlations between all conditions were informative, moderate and positive (estimates > 0.43), reflecting baseline variability in sensitivity across participants. Additionally, numerical trends favored higher baseline variability in either production condition relative to the silent condition.

Item-level random slopes corresponding to the effects of condition and group were informative, positive, and similar across groups (estimates > 0.34 for the matched group and > 0.37 for the unmatched group). This is consistent with variability in the impact of our fixed effects across items. Like our other analyses, item-level correlation terms were less informative albeit positive in all cases, suggesting some degree of baseline variation in sensitivity across items, albeit less variability relative to our other models.

*Response Bias.* As depicted in Figure 2, we again observed credibly lower bias for either production condition relative to the silent condition. For this dependent measure, however, the difference between the sing and aloud conditions was centered on zero (difference = -0.01, HDI95% = -0.15 – 0.13).

For our analysis of recollection, participant-level random slopes corresponding to the effect of condition were informative in all cases (all estimates > 0.54), indicating variability in the impact that study modality had on response bias across individuals. Participant-level correlations between all conditions were strong and positive (estimates > 0.80), reflecting baseline variability in response bias across participants.

Item-level random slopes corresponding to the effects of condition and group were also informative in all cases and similar across groups (estimates > 0.22 for the matched group and > 0.24 for the unmatched group). For response bias, item-level correlations were less informative but positive in all cases (estimates > 0.17 for the matched group and > 0.30 for the unmatched group). For this parameter, numerical trends favored a greater degree of baseline variability across items in the unmatched group relative to the matched group.

**Familiarity**

Finally, we analyzed familiarity, which is often viewed as a nonspecific feeling of fluency or familiarity that can drive recognition responses (Yonelinas, 2002). Familiar responses from recollect/familiar/neither judgements were binarized such that trials for which participants responded with “F” indicated a “familiar” response. However, estimating familiarity using the raw trial data for which a familiar response was made is liable to underestimate the parameter: Trials in which participants indicate recollection likely still involve some degree of familiarity, but the former response takes precedence over the latter. To mitigate this problem, some theorists have advocated for the use of the *Independence Remember-Know* *Procedure*, which produces estimates of familiarity that better align with other techniques used to estimate the parameter (see, e.g., Yonelinas, 2002; Yonelinas & Jacoby, 1995). Typically, this procedure involves dividing the raw proportion of trials for which “familiar” responses were made by the raw proportion of trials for which “recollect” responses were not made. For the purposes of our analyses, however, we opted instead to apply a probit regression to trial data for which “recollect” response were not made. Estimating familiarity using this methodology is equivalent to conventional calculations of the Independence Remember-Know Procedure (for further discussion and mathematical proof, see Fawcett et al., 2016; see also, Fawcett & Ozubko, 2016). Aside from these differences, the model we applied to “familiar” responses was otherwise identical to that described for the previous analyses. The estimates reported hereafter can be interpreted much like those reported for confidence ratings and recollection, albeit with estimates for sensitivity representing participants’ propensity to successfully discriminate between old and new items with “familiar” responses.

*Sensitivity.* As depicted in Figure 1, analysis of the familiarity responses followed the same pattern observed for our other dependent measures, with credible production effects for either modality. 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. However, we once again observed little support for an SSE or a difference in the magnitude of the SSE between matching conditions, difference = 0.18 (HDI95% = -0.14 – 0.49).

For our analysis of familiarity, participant-level random slopes corresponding to the effect of condition were informative (all estimates > 0.10), albeit numerically smaller than those observed for either confidence or recollection. Nonetheless, this pattern supports some degree of variability in the impact that study modality had on sensitivity across individuals. Interestingly, numerical trends favored greater variability in sensitivity for the sing condition relative to either the silent or aloud conditions. Participant-level correlations between all conditions were less informative, albeit positive (estimates > 0.43), reflecting some degree of baseline variability in sensitivity across participants.

Item-level random slopes corresponding to the effects of condition and group were informative, positive, and similar across groups (estimates > 0.15 for the matched group and > 0.16 for the unmatched group), consistent with variability in the impact of our fixed effects on familiarity sensitivity across items. Like our analysis of confidence ratings, item-level correlation terms were less informative albeit positive in all cases, suggesting some degree of baseline variation in sensitivity across items.

*Response Bias.* As depicted in Figure 2, we observed credibly lower bias for aloud items relative to silent items. Although the difference between the sing and silent conditions was not credible, a numerical trend favored lower response bias for the former. For familiarity, however, we observed credibly higher bias for the sing condition relative to the aloud condition (difference = 0.29, HDI95% = 0.08 – 0.33). This is generally consistent with the patterns observed for confidence ratings.

For our analysis of familiarity, participant-level random slopes corresponding to the effect of condition were informative in all cases (all estimates > 0.65), indicating variability in the impact that study modality had on response bias across individuals. Participant-level correlations between all conditions were strong and positive (estimates > 0.86), reflecting baseline variability in response bias across participants.

Item-level random slopes corresponding to the effects of condition and group were also informative in all cases and similar across groups (estimates > 0.15 for the matched group and > 0.16 for the unmatched group). In either group, numerical trends favored higher variability for either production condition relative to the silent condition. For this model, item-level correlations were informative, moderate, and positive in all cases (estimates > 0.48 for the matched group and > 0.41 for the unmatched group); these terms were similar across groups. Thus, much like our other analyses, this model suggests that there is baseline variability in response bias across items.

**Meta-Analysis of the Singing Superiority Effect**

Having failed to replicate the SSE, we opted to conduct a meta-analysis of the extant literature on this topic to provide a stronger empirical test of this phenomenon.

**Method**

**Search and Coding**

Our meta-analysis made use of a recent search already conducted of the production effect literature (Fawcett et al., 2022). While those authors were concerned with the between-subject production effect, this search used the broad keyword “production effect” with no modifiers and should thereby have captured the vast majority of relevant literature. It also included all articles citing or cited by key articles in the area (i.e., Fawcett, 2013; MacLeod et al., 2010) and forward and backward snowball searches of articles included in their search. For the present investigation, the search conducted by Fawcett et al. (2022) was combined with forward and backward snowball searches of each of our included studies.

Of the articles reviewed using our combined search methods, three were identified for inclusion based on their reporting a within-subject production manipulation including both sing and aloud conditions: Quinlan and Taylor (2013, 2019) and Hassall et al. (2016). Including the experiment reported in the present study, our sample consisted of 13 independent experiments reported across four studies. Means, standard deviations, sample sizes and correlations corresponding to comparisons between conditions were recorded for each included experiment.

**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 the sing and aloud 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 calculating by aggregating hits and false alarm rates into proportions and applying transformations to the data (see, e.g., Stanislav & Todorov, 1999). Because estimates of variability for differences between conditions were not available for Quinlan and Taylor (2013, Experiment 3), we imputed this parameter using the other data available to us.[[4]](#footnote-4)

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 parameters corresponding to both aggregate effects and between-study heterogeneity, particularly in cases where the sample of effects being aggregated is small (e.g., Harrer et al., 2021; Williams et al., 2018).[[5]](#footnote-5) 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 (for further discussion of the advantages of Bayesian credible intervals over Frequentist confidence intervals, see, e.g., Morey et al., 2016).

We modeled our data such that our dependent measure was the raw mean difference in *d’* (calculated conventionally; see, e.g., Stanislaw & Todorov, 1999) between the sing and aloud conditions. The parameterization was analogous to a Frequentist random effects meta-analysis such that our models estimated the size of the singing superiority effect weighted by sampling variance and included a random intercept corresponding to the experiment from which the effect was derived. Thus, our models assumed variability in the size of the singing superiority effect across experiments; given heterogeneity across samples, sites, and methodologies used in investigations of the effect, we believe this assumption is justified (for further discussion, see, e.g., Borenstein et al., 2010). With this parameterization in mind, our models computed an intercept corresponding to the estimated aggregate singing superiority effect across studies. Additionally, our computed a random intercept corresponding to between-study heterogeneity, akin to tau in Frequentist models (see, e.g., Harrer et al., 2021).

We also applied uninformative, mildly regularizing priors to our meta-analytic models. These priors reflected our belief that the size of the raw mean difference in sensitivity should reasonably fall between -0.6 and 0.6 in a typical study, with effects in individual studies permitted to range from -1.2 to 1.2. These priors were calibrated with respect to previous effects reported by Quinlan and Taylor (2013, 2019; Hassall et al., 2016), who observed raw mean differences in *d’* scores ranging from ~ 0.1 to 0.5. Additionally, these priors reflect our *a priori* theoretical belief that a slightly more elaborate type of vocalization should not be vastly superior to an already large benefit; nonetheless, our priors did allow for very large singing superiority effects in individual studies.

All meta-analytic models were fit using 4 independent chains of 80000 iterations each with a warm-up period of 40000 iterations. Model convergence was assessed using R-hat statistics, which were less than 1.00 in all cases, indicating that all models converged (Gelman & Hill, 2006; Kruschke, 2010). Further, inspection of the chains showed that effective sample size was greater than 70000 for all estimates.

**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).

As depicted in Figure 3, the aggregate SSE was credible, with the difference between the sing and aloud conditions estimated at 0.16 (HDI95% = 0.05 – 0.26). Although the aggregate estimate was 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 -0.07 to 0.42; this implies that some studies show roughly no effect whereas others show effects that are quite large. The random intercept permitting baseline variability in the size of the SSE was also informative (estimate = 0.12, HDI95% = 0.01 – 0.24), supporting the notion that between-study heterogeneity in the size of the SSE. 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 that the SSE – if truly reliable – has likely been overestimated.

Given that our experiment hinted at the possibility that foil matching might play an important role in facilitating the SSE, we conducted an exploratory meta-analysis that included colour matching as a moderator. This model was parameterized similarly to that described above, albeit with the inclusion of a categorical fixed effect corresponding to whether each experiment used colour matching. Further, because our model now included a categorical fixed effect, we removed the model intercept and computed slopes corresponding to the aggregate SSEs for colour matched and unmatched studies, respectively (akin to a subgroup analysis; see, e.g., Borenstein & Higgins, 2013). Here, the aggregate SSE was credible when colour matching was present, with the difference between the sing and aloud conditions estimated at 0.23 (HDI95% = 0.11 – 0.34; PI95% = 0.02 – 0.43). However, the effect was not credible in the absence of this procedure, estimated at 0.05 (HDI95% = -0.08 – 0.18; PI95% = -0.19 – 0.28). Interestingly, the random intercept corresponding to between-study heterogeneity was numerically smaller and less informative relative to that observed in our previous model (estimate = 0.08, HDI95% = 0.00 – 0.20), suggesting that less between-study heterogeneity was present in the model after accounting for the use of colour matching. Consistent with the patterns we observed in our experiment, these results suggest that the SSE might emerge only when colour matching is used at test.

Finally, to evaluate publication bias, we first fit two multilevel models analogous to Egger’s regression test; this procedure can be used to evaluate associations between effect sizes and the precision with which the effects were estimated (Egger et al., 1997). In either case, priors, sampling procedures and random effects structures were identical to our other meta-analyses of the singing superiority effect. These models were parameterized similarly to conventional calculations of Egger’s regression test, albeit inclusive of our random effects; the random effect structure of these models and the corresponding assumptions were identical to that reported above. Each model estimated the size of the singing superiority effect weighted by sampling variance and included a fixed effect for either standardized sample size or standard error. Thus, each model respectively computed an intercept reflecting the aggregate SSE for a study with an average sample size or standard error. Additionally, each model computed a random intercept reflecting between-study heterogeneity (described above) and a slope corresponding to the included fixed effect. For the model including sample size, the intercept was estimated at 0.16 (HDI95% = 0.07 – 0.26) and the slope for sample size was not credible (estimate = -0.09, HDI95% = -0.19 – 0.02), providing no credible evidence for publication bias. For the model including standard error, the intercept was estimated at 0.16 (HDI95% = 0.05 – 0.27) and the slope for standard error was not credible (estimate = 0.03, HDI95% = -0.09 – 0.14), again providing no credible evidence for publication bias. For either model, the estimates corresponding to the random intercept were informative and comparable in size to that of our first meta-analytic model.

Subsequently, we fit a cumulative meta-analysis of the SSE, wherein studies were added to the model iteratively in order of sample size (largest to smallest; see, e.g., Leimu & Koricheva, 2004). To accomplish this, we began by fitting a model of the singing superiority effect that was parameterized identically to our first meta-analytic model reported above, albeit including only the largest study. Subsequently, we added the next largest study and re-fit the model; this procedure was repeated until all studies had been included. For all iterative models, our priors, sampling procedures and random effects structures were identical to our other meta-analyses of the singing superiority effect. As shown in Figure 4, the aggregate SSE was small and non-credible when only large studies were included. The aggregate estimate was credible only after small studies (*N* < 24 participants) were added to the model; this pattern is consistent with an aggregate SSE that is driven predominantly by small sample effects. However, it could also be that sample size is correlated with colour matching (which was the only condition to show a credible effect).

**General Discussion**

The present study evaluated evidence favouring singing as a mnemonically superior production modality compared to reading aloud. Across a conceptual replication of previous work on the singing superiority effect and a meta-analysis of all published studies – we found the SSE to be smaller than previously thought, and possibly dependent on provision of the study phase conditions at test. In our experiment, we replicated the methods of Quinlan and Taylor (2013) more closely than previous efforts in order to account for the potential role that colour matching might play in facilitating the SSE. Here, we observed robust production effects for both reading aloud and singing across all our dependent measures, but a credible SSE failed to emerge. Our meta-analysis of all known studies demonstrated evidence of a small SSE, but moderator analysis revealed this effect to be dependent on the test items being presented in the colour corresponding to their study phase condition.

Based on these findings, support for the sensorimotor scaling hypothesis would appear to be limited. At the 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 difference in sensitivity of ~ 0.36, whereas our meta-analytic estimate was 0.16 overall and 0.23 in studies using colour 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 – 43) reported smaller effect sizes better aligned with the differences observed in the present investigation (e.g., *MD* = ~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* = ~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 colour 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; on the other hand, knowing a test item was sung or read aloud might encourage reactivation or utilization 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 colour 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 explanation is that colour 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 test as a means of discriminating between items (e.g., by thinking about saying the item aloud; see also, Zhou, 2022). 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*, it is recreating the productive act in one’s mind that offers the additive advantage of singing (or other forms of particularly elaborate modes of speaking, such as the character voices as used by Wakeham-Lewis et al., 2022). Providing cues about how an item would have been produced might guide participants to reinstate production in a manner more 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 colour 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.[[6]](#footnote-6)

Perhaps the simplest explanation 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 (Xu, 2005), pitch, (Bent et al., 2006; Dolson, 1994) and timbre (Terasawa et al., 2005). 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 quantitative. Rather than appending additional features to the production trace, then, singing might instead allow for a greater degree of variation in item representations across articulatory and auditory features. Such a model might also accommodate an interaction between singing and colour matching at test: Regardless of modality, produced items share common 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 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 experiment (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 detect an effect across our dependent measures. Our experiment 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 were supervised throughout the study phase and prompted participants to sing more effortfully if their singing faltered; given that 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. While Hassall et al. (2016; see also, 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.

In sum, the present investigation poses a challenge to the SSE as described in earlier literature (e.g., Quinlan & Taylor, 2013, 2019). Across our analyses, we observed a production effect for singing that was similar in magnitude to that for reading aloud (see also, Hassall et al., 2016). When the SSE did emerge, it was much smaller than previous estimates and was confined to the colour matched group. Contrary to sensorimotor scaling explanations of the effect, then, it appears that the relative superiority of singing arises on the basis of factors related to study design. 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.

**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>

Banks, W. P. (1970). Signal detection theory and human memory. *Psychological Bulletin, 74*(2), 81–99. [https://doi.org/10.1037/h0029531](https://psycnet.apa.org/doi/10.1037/h0029531)

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language, 68*(3), 255–278. [https://doi.org/10.1016/j.jml.2012.11.001](https://psycnet.apa.org/doi/10.1016/j.jml.2012.11.001)

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](https://psycnet.apa.org/doi/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*, 149-154.<https://doi.org/10.3758/s13423-013-0485-1>

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2010). A basic introduction to fixed-effect and random-effects models for meta-analysis. *Research Synthesis Methods, 1*(2), 97–111. [https://doi.org/10.1002/jrsm.12](https://psycnet.apa.org/doi/10.1002/jrsm.12)

Broadbent, D. E. (1967). Word-frequency effect and response bias. *Psychological Review, 74*, 1–15. [https://doi.org/10.1037/h0024206](https://psycnet.apa.org/doi/10.1037/h0024206)

Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods, 41,* 977–990. <https://doi.org/10.3758/BRM.41.4.977>

Bürkner, P. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*, 1–28. <https://doi.org/10.18637/jss.v080.i01>

Coltheart, M. (1981). The MRC psycholinguistic database. *The Quarterly Journal of Experimental Psychology: Human Experimental Psychology, 33*(4), 497–505. <https://doi.org/10.1080/14640748108400805>

DeCarlo, L. T. (1998). Signal detection theory and generalized linear models. *Psychological Methods, 3*(2), 186–205. [https://doi.org/10.1037/1082-989X.3.2.186](https://psycnet.apa.org/doi/10.1037/1082-989X.3.2.186)

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>

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://psycnet.apa.org/doi/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](https://psycnet.apa.org/doi/10.1037/xlm0001093)

DeCarlo, L. T. (1998). Signal detection theory and generalized linear models. *Psychological Methods, 3*(2), 186–205. [https://doi.org/10.1037/1082-989X.3.2.186](https://psycnet.apa.org/doi/10.1037/1082-989X.3.2.186)

Dodson, C. S., & Schacter, D. L. (2001). If I had said it I would have remembered it: Reducing false memories with a distinctiveness heuristic. *Psychonomic Bulletin & Review, 8*, 155–161. <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>

Egan, J. P. (1958). Recognition memory and the operating characteristic. *USAF Operational Applications Laboratory Technical Note, 58-51,* ii, 32.

Egger, M., Smith, G. D., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *BMJ*, *315*(7109), 629–634. <https://doi.org/10.1136/bmj.315.7109.629>

Fawcett, J. M. (2013). The production effect benefits performance in between-subject designs: A meta-analysis. *Acta Psychologica, 142*, 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](https://psycnet.apa.org/doi/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>

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, 1046–1055.<https://doi.org/10.3758/s13421-012-0210-8>

Furukawa, T. A., Barbui, C., Cipriani, A., Brambilla, P., & Watanabe, N. (2006). Imputing missing standard deviations in meta-analyses can provide accurate results. *Journal of Clinical Epidemiology*, *59*, 7–10. <https://doi.org/10.1016/j.jclinepi.2005.06.006>

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>

Gelman, A., & Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press.

Grady, C. L., McIntosh, A. R., Horwitz, B., Maisog, J. M., Ungerleider, L. G., Mentis, M. J., Pietrini, P., Schapiro, M. B., & Haxby, J. V. (1995). Age-related reductions in human recognition memory due to impaired encoding. *Science, 269*(5221), 218–221. [https://doi.org/10.1126/science.7618082](https://psycnet.apa.org/doi/10.1126/science.7618082)

Harrer, M., Cuijpers, P., Furukawa, T.A., & Ebert, D.D. (2021). *Doing Meta-Analysis with R: A Hands-On Guide*. Boca Raton, FL and London: Chapmann & Hall/CRC Press.

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>

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), 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>

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*, 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>

Leimu, R., & Koricheva, J. (2004). Cumulative meta-analysis: A new tool for detection of temporal trends and publication bias in ecology. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, *271*(1551), 1961-1966. <https://doi.org/10.1098/rspb.2004.2828>

Lewandowski, D., Kurowicka, D., & Joe, H. (2009). Generating random correlation matrices based on vines and extended onion method. *Journal of Multivariate Analysis*, *100*(9), 1989–2001. <https://doi.org/10.1016/j.jmva.2009.04.008>

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>

Macmillan, N. A., & Creelman, C. D. (2005). *Detection theory: A user's guide* (2nd ed.). Lawrence Erlbaum Associates Publishers.

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*, 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*, 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](https://psycnet.apa.org/doi/10.3758/s13423-020-01762-3)

McElreath, R. (2018). *Statistical rethinking: A Bayesian course with examples in R and Stan*. Chapman and Hall/CRC.

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*, 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>

Paivio, A., Walsh, M., & Bons, T. (1994). Concreteness effects on memory: When and why? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20*(5), 1196–1204. [https://doi.org/10.1037/0278-7393.20.5.1196](https://psycnet.apa.org/doi/10.1037/0278-7393.20.5.1196)

Paivio, A., Yuille, J. C., & Madigan, S. A. (1968). Concreteness, imagery, and meaningfulness values for 925 nouns. *Journal of Experimental Psychology, 76*, 1–25. [https://doi.org/10.1037/h0025327](https://psycnet.apa.org/doi/10.1037/h0025327)

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*. <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, Vienna, Austria. Retrieved from <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*, 10–25. <https://doi.org/10.1016/j.jml.2012.09.001>

Rouder, J. N., & Lu, J. (2005). An introduction to Bayesian hierarchical models with an application in the theory of signal detection. *Psychonomic Bulletin & Review*, *12*(4), 573-604. <https://doi.org/10.3758/BF03196750>

Rouder, J. N., Lu, J., Sun, D., Speckman, P., Morey, R., & Naveh-Benjamin, M. (2007). Signal detection models with random participant and item effects. *Psychometrika*, *72*(4), 621-642. <https://doi.org/10.1007/s11336-005-1350-6>

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*, 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](https://psycnet.apa.org/doi/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>

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>

Terasawa, H., Slaney, M., & Berger, J. (2005). A timbre space for speech. *Interspeech, 2005*, 1729–1732.

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](https://psycnet.apa.org/doi/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>

Vuorre, M. (2017, October 9). Bayesian estimation of signal detection models. <https://mvuorre.github.io/posts/2017-10-09-bayesian-estimation-of-signal-detection-theory-models/>

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](https://psycnet.apa.org/doi/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>

Wright, D. B., Horry, R., & Skagerberg, E. M. (2009). Functions for traditional and multilevel approaches to signal detection theory. *Behavior Research Methods*, *41*, 257–267. <https://doi.org/10.3758/BRM.41.2.257>

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>

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>

Zormpa, E., Brehm, L. E., Hoedemaker, R. S., & Meyer, A. S. (2019). The production effect and the generation effect improve memory in picture naming. *Memory*, *27*(3), 340–352. <https://doi.org/10.1080/09658211.2018.1510966>

**Figure 1**

*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)*

A graph of a graph showing the number of individuals

Description automatically generated with medium confidence

*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.

**Figure 2**

*Posterior Estimates for Response Bias (C) as a Function of Condition and Group (Left Column) and Contrasts Between Conditions as a Function of Group (Right Column)*

A graph of a number of individuals

Description automatically generated with medium confidence

*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.

**Figure 3**

*Forest Plot Depicting Raw Mean Differences in d’ (Sing-Aloud) for a Meta-Analytic Model of the Singing Superiority Effect*

A graph of a graph showing the size of a number of objects

Description automatically generated with medium confidence

*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 4**

*Forest Plot Depicting Raw Mean Differences in d’ (Sing-Aloud) for a Cumulative Meta-Analytic Model of the Singing Superiority Effect*

A graph of a number of numbers

Description automatically generated with medium confidence

*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. Studies were added in order of sample size, starting with the largest study (at the top) and adding one study at a time until all studies were included (at the bottom).

1. We recorded confidence ratings because it had been our intention to analyze these and the data to follow using a multilevel ordinal regression model within. However, since conducting these studies, we have become aware that this is not yet possible in the manner we had intended owing to limitations of the *brms* package (Bürkner, 2017) with respect to random effects for thresholds. For that reason, and because the gains from conducting ordinal as opposed to binarized probit models are modest, we have instead adopted a more traditional approach throughout. [↑](#footnote-ref-1)
2. Our models used the traditional dummy coding of 0.5 and -0.5 for estimates of *C* (Macmillan & Creelman, 2005), which produces analogous estimates on a more conventional scale. [↑](#footnote-ref-2)
3. Modeling correlations between item- and participant-level random slopes accounts for the notion that some participants or items may vary in baseline sensitivity or response bias. For example, if participant-level slopes for sensitivity are positively correlated across conditions, this indicates that participants who exhibit high sensitivity in one condition also tend to exhibit high sensitivity for other conditions. [↑](#footnote-ref-3)
4. In this case, we imputed the missing standard deviation using average of the parameter across our other experiments (see, e.g., Furukawa et al., 2006). [↑](#footnote-ref-4)
5. The superiority of Bayesian estimates arises in part because Bayesian approaches to meta-analysis incorporate uncertainty into estimates of between-study heterogeneity, whereas Frequentist approaches do not (Harrer et al., 2021). [↑](#footnote-ref-5)
6. Alternatively, it could be the case that the SSE is simply too small to detect reliably across experiments. However, our findings argue against this notion: Our experiment failed to detect a credible effect despite using samples that were much larger – and thereby better powered – than those used in previous efforts. [↑](#footnote-ref-6)