DOES THE SONG REMAIN THE SAME? FURTHER INVESTIGATIONS OF THE SINGING SUPERIORITY EFFECT

by © Jedidiah William Whitridge

Thesis submitted to the School of Graduate Studies

in partial fulfillment of the requirements for the degree of

Master of Science in Experimental Psychology

Department of Psychology

Faculty of Science

Memorial University of Newfoundland

August 2024

St. John's, Newfoundland and Labrador

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 than reading aloud (the *singing superiority effect*), possibly due to the inclusion of further sensorimotor features (e.g., more variable tone). However, the singing superiority effect has not always replicated. Across two experiments, I demonstrated robust production effects for both reading aloud and singing but observed a singing superiority effect only when items were tested in the same colour in which they were studied (with foils randomized to colour). 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

General Summary

People tend to exhibit better memory for words that they read aloud (i.e., produce) relative to words they read silently; this phenomenon is known as the *production effect*. This finding is thought to occur because the additional sensory processing that occurs when producing words (e.g., moving one's mouth, hearing oneself say the word) renders the words *distinctive* in memory relative to those read silently. One piece of evidence supporting this hypothesis is the finding that singing words results in an especially large benefit to memory that is superior even to reading aloud. This *singing superiority effect* has been explained by some as occurring due to additional sensory processing that is unique to singing. However, the present thesis provides evidence that the relative superiority of singing over reading aloud has been overstated, and that the singing superiority effect is likely driven by idiosyncratic methodological factors rather than additional sensory processing. These findings pose a challenge to dominant, distinctiveness-based theories of the production effect.

Acknowledgements

I would like to thank my supervisor, Dr. Jonathan Fawcett, for his guidance and support throughout my program of study. Additionally, I would like to thank our collaborators, without whom this project would not have been possible.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

Of the faculties possessed by the human mind, few are as remarkable or important as our capacity to remember. Without memory, we would be ill-equipped to function altogether; in fact, some researchers have gone as far as to suggest that our conceptions of self are constructed and defined by our memories (e.g., Robinson & Taylor, 2014). Given our reliance upon this faculty, it is unsurprising that there exists long-standing interest – academic and otherwise – in how we might improve our ability to remember information (e.g., James, 1890). Perhaps one of the oldest and most intuitive mnemonic strategies is the act of reading to-be-remembered information aloud. Historically, this strategy has been employed by ancient Greeks to aid memorization of poetry (e.g., Lentz, 1985) and by scholars in the Middle Ages as a means of committing religious texts to memory (e.g., Burke & Ornstein, 2018). In more recent contexts, college students commonly report reading lecture notes and textbook passages aloud as a method of learning course material and studying for examinations (e.g., Morehead et al., 2016). However, despite millennia of usage and intuitive wisdom favoring the benefits of reading information aloud, formal academic study of this strategy began only half a century ago (e.g., Crowder, 1970; Hopkins & Edwards, 1972; Kappel et al., 1973; Routh, 1970). Since the inception of this field of research, empirical findings have converged in support of the notion that information read aloud is remembered better than information read silently, a phenomenon that has since been termed the *production effect* (MacLeod et al., 2010).

The earliest documentation of the production effect is commonly attributed to Hopkins and Edwards (1972), who demonstrated that participants remember words read aloud significantly better than those read silently using a mixed-list recognition paradigm.[1] The

[1] Although Hopkins and Edwards (1972) were apparently the first to document the production effect using words as stimuli, earlier reports of mnemonic advantages for produced stimuli using digit span paradigms exist (e.g., Crowder, 1970; Routh, 1970).

phenomenon remained relatively obscure in the years that followed, although sporadic investigations during this period extended the production effect across modalities (e.g., writing, mouthing) and began to develop theoretical frameworks to explain the benefit (e.g., Conway & Gathercole, 1987; Dodson & Schacter, 2001; Gathercole & Conway, 1988; Greene & Crowder, 1984; Kappel et al., 1973; MacDonald & MacLeod, 1998). Nearly four decades after the phenomenon was first documented, MacLeod et al. (2010) reignited academic interest in the production effect, widening its boundaries and generating new theoretical knowledge. The effect has since been shown to persist across a variety of production modalities, with benefits demonstrated for spelling, typing (Forrin et al., 2012), drawing (Namias et al., 2022; Wammes et al., 2016), signing (Taitelbaum-Swead et al., 2018) and even imagining the act of production (Jamieson & Spear, 2014). Further, whilst typical studies of the production effect tend to employ mixed-list, recognition paradigms, research has demonstrated the benefit to be robust across variations in experimental design (e.g., free recall; Lin & MacLeod, 2012; between-subjects or pure-list manipulations; Bodner et al., 2014; for a review, see Fawcett et al., 2023), stimuli (e.g., sentences; Ozubko et al., 2012a; pictures; Fawcett et al., 2012), and populations (e.g., older adults; Icht et al., 2022; Lin & MacLeod, 2012; young children; Icht & Mama, 2015; Pritchard et al., 2020). Finally, research in the applied domain has explored how the production effect might be leveraged to improve verbal learning in clinical populations (e.g., Icht et al., 2019; Mama & Icht, 2019) or as a study aid (e.g., Ozubko et al., 2012a).

Although practical applications of the production effect show promise, the majority of research on the phenomenon has been theoretically – rather than practically – motivated. Early investigations speculated that a production-related benefit might arise due to deeper processing of produced items (e.g., Crowder, 1970), additional sensory information present at encoding

(e.g., Kappel et al., 1973), or impaired encoding of unproduced items (e.g., Hopkins & Edwards, 1972; Routh, 1970). More recently, the latter two explanations have been refined into the *distinctiveness account*, which suggests that production (relative to reading) renders stimuli distinctive in memory, facilitating better retrieval at test (Conway & Gathercole, 1987; Dodson & Schacter, 2001; MacLeod et al., 2010). This framework has garnered a great deal of empirical support (e.g., MacLeod et al., 2010; Ozubko & MacLeod, 2010; Richler et al., 2013) and is generally accepted as the dominant theoretical account of the production effect (Fawcett, 2013; Fawcett et al., 2023; MacLeod & Bodner, 2017).

However, other theorists have challenged explanations of the production effect that rely solely on distinctiveness, proposing instead roles for strength of encoding (e.g., Bodner et al., 2016; Fawcett & Ozubko, 2016; Bodner & Taikh, 2012), differential retention of item-order information (e.g., Jonker et al., 2014; Lambert et al., 2016), and variation in attentional processes at encoding (e.g., Fawcett et al., 2022; Mama et al., 2018; Willoughby et al., 2019). One theoretical challenge to the distinctiveness account has arisen on the basis of findings that the magnitude of the production effect does not necessarily increase as the productive act becomes more distinctive (e.g., Hassall et al., 2016; Wakeham-Lewis et al., 2022; Whitridge, 2022). The present thesis explored this finding further by manipulating singing as a study modality in production tasks (Hassall et al., 2016; Quinlan & Taylor, 2013, 2019; see also, Forrin et al., 2012). In the sections that follow, I review evidence pertaining to both the distinctiveness account and alternative theories of the production effect. Subsequently, I discuss potential theoretical bases of the production effect for singing in relation to distinctiveness frameworks.

1.1 Distinctiveness Accounts of the Production Effect

Modern distinctiveness-based accounts of the production effect derive largely from two early assertions: (1) For produced items, additional sensory processing occurs at encoding, a record of which exists in memory and can be accessed later (Kappel et al., 1973); and (2) the benefit for produced items can occur only *relative* to unproduced items (Hopkins & Edwards, 1972). In their present incarnation, distinctiveness-based accounts generally contend that producing an item appends additional cues (or *features*) to the memory trace associated with the item. These features represent a record of sensorimotor processing that occurs at encoding for produced (but not unproduced) items (e.g., auditory and motoric processing); the resulting association between produced items and processing records has been referred to by some as the *production trace* (Fawcett, 2013; Fawcett et al., 2012). At test, items associated with distinctive features are thought to stand out against the backdrop of unproduced items, yielding an advantage for the former (Forrin et al., 2012; MacLeod et al., 2010; see also Conway & Gathercole, 1987; Dodson & Schacter, 2001).

There exists heterogeneity within distinctiveness-based accounts, with some variants contending that the production trace is leveraged consciously in the form of a *distinctiveness heuristic*: Participants are thought to scan their episodic memory for information about having produced an item at study and use this information to guide discrimination between old and new items on recognition tests ("I remember saying this item aloud, so I must have studied it;" Dodson & Schacter, 2001; MacLeod et al., 2010). On the other hand, frameworks derived from feature-based models of human memory (e.g., MINERVA 2; Hintzman, 1984) suggest that the presence of distinctive features might intrinsically render items more easily retrievable: If retrieval is dependent upon activation of features, items with a larger number of associated

features (e.g., produced items) possess an inherent advantage, eliminating the need for distinctiveness to be employed strategically (for a detailed discussion, see Jamieson et al., 2016).

Regardless of whether distinctiveness is leveraged consciously or otherwise, the framework has been well-supported empirically as an explanation for the production effect. One key piece of evidence for this account is the finding that manipulating the utility of distinctiveness can reduce or eliminate the production effect (e.g., Bodner et al., 2016; Icht et al., 2014; MacLeod et al., 2010; Ozubko & MacLeod, 2010; Richler et al., 2013). For example, MacLeod et al. (2010, Experiment 4) had participants study a mixed list wherein "aloud" items were produced either by pressing a key or saying the word "yes" aloud; importantly, participants in a given condition produced each "aloud" item in the same manner. Under these conditions, the production effect was eliminated altogether, with similar recognition performance for produced and unproduced items. Although additional sensorimotor processing would be expected to occur at study for both keypresses (i.e., motoric processing) and "yes" responses (i.e., motoric and auditory processing), the record of processing was common across produced items in this paradigm, diminishing the utility of the production trace (see also, Castel et al., 2013).

These findings were later replicated and extended by Richler et al. (2013) in a picture naming paradigm: Relative to silent naming, producing the full name of an object yielded a significant advantage in recognition memory, whereas using a non-distinct keypress to label an object did not.[2] Richler et al. further observed that producing object names led to a benefit only when the names of objects were distinctive; labelling aloud object exemplars that belonged to a common category (e.g., *chair*, *lamp*, etc.) yielded no advantage in discrimination relative to

[2] Although Richler et al. (2013) has been accepted as evidence for the distinctiveness account (see, e.g., MacLeod & Bodner, 2017), it is likely that the influence of production in this study was confounded that of response generation, which may impact interpretation of the results (see Whitridge et al., submitted; Zormpa et al., 2019).

silent naming. Taken together, the findings of MacLeod et al. (2010) and Richler et al. (2013) suggest that the mere presence of additional sensorimotor information is not adequate to produce a mnemonic benefit for produced items. Rather, the production trace must be sufficiently item-specific for a production effect to occur, as distinctiveness accounts predict.

Similarly, decreasing the diagnostic value of the production trace appears to impede the production effect. Experiments by Ozubko and MacLeod (2010) utilized a source monitoring paradigm wherein participants studied two lists – a pure list (i.e., a list wherein all items are studied either aloud or silently) and a mixed list (i.e., a list wherein some items are studied aloud and some are studied silently) – and later completed a recognition test that required judgements as to which list each item originated from. The authors found that the production effect in source monitoring (see Ozubko et al., 2012b, 2014) was eliminated when participants studied the pure list aloud, but not silently. Because participants in the pure-aloud condition produced words from both lists, Ozubko and MacLeod (2010) suggested that participants' records of having produced a word were not distinctive to either list; the production trace would thereby have possessed little diagnostic value in discriminating between sources. Conversely, when silent pure lists were studied, participants could successfully leverage distinctive information about having studied words aloud in only one of the lists, leading to the production benefit that is typically observed in source monitoring (e.g., Ozubko et al., 2012b, 2014). In support of a distinctiveness account, then, these results indicate that even an item-specific record of processing is not diagnostic unless distinctive information can be leveraged to guide discrimination.

However, not all empirical evidence favors distinctiveness-based accounts. One early prediction made by this framework was that the production effect should be confined to mixed list (i.e., within-subject) designs and should not occur for pure list (i.e., between-subject) designs

(Hopkins & Edwards, 1972; MacLeod et al., 2010); according to Hunt (2006), distinctiveness is not absolute and can emerge only when some stimuli stand out relative to others. Although a number of early investigations failed to observe a benefit when production was manipulated between-subject (e.g., Dodson & Schacter, 2001; Hopkins & Edwards, 1972; MacLeod et al., 2010), such an effect has since been observed in experimental studies (e.g., Bodner et al., 2014; Bodner et al., 2016; Forrin et al., 2016; Taikh & Bodner, 2016; see also, Gathercole & Conway, 1988; Greene & Crowder, 1984; Kappel et al., 1973) and supported further by meta-analyses (Fawcett, 2013; Fawcett et al., 2023; see also, Bodner et al., 2014). Recent meta-analytic efforts by Fawcett et al. (2023) suggested the apparent unreliability of the between-subject production effect in previous investigations could be attributed largely to underpowered experiments: The benefit is smaller in between-subject designs and therefore requires larger samples to detect reliably.

Initially, the between-subject production effect was thought to occur only in tests of recognition, with no reliable advantage being observed in recall (e.g., Jones and Pyc, 2014; but see Greene & Crowder, 1984; Kappel et al., 1973).[3] However, recent experiments by Saint-Aubin and colleagues (Cyr et al., 2022; Saint-Aubin et al., 2021; Gionet et al., 2022) have revealed that production interacts with serial position in between-subject recall paradigms: While a reliable production effect occurs for items in later serial positions, a reverse production effect (i.e., silent > aloud) occurs for early positions (see Fawcett et al., 2023 for a meta-analysis of this effect). These opposing effects seemingly "cancel out," such that recall performance for whole lists of aloud items is similar to that of silent items; this mechanism may have obfuscated the between-subject production effect in recall paradigms that failed to observe such a benefit (e.g.,

[3] Although production does not necessarily appear to benefit item memory in between-subject recall paradigms, recent meta-analytic efforts have demonstrated a reliable, production-related reduction in intrusions for such paradigms (Fawcett et al., 2023).

Jones & Pyc, 2014). This unusual pattern of results is believed to occur because producing items interferes with rehearsal. Typically, primacy effects in serial recall occur because early items benefit from rehearsal to a greater extent than later items (e.g., Marshall & Werder, 1972). In production paradigms, however, interference due to sensorimotor processing of subsequent items eliminates both the primacy advantage and the production effect for early items. On the other hand, the benefit of production persists for later items because they are not subject to the same degree of sensorimotor interference (i.e., because few or no additional items are produced prior to the test). Although the task-specific demands of recall paradigms result in a nuanced pattern of findings, these studies nonetheless demonstrate that the between-subject production effect – much like the within-subject effect – is robust to changes in experimental design.

Considered in aggregate, evidence for a between-subject production effect poses a substantial theoretical challenge to distinctiveness-based frameworks of the effect. Interestingly, however, some authors suggest that the existence of a smaller between-subject production effect may in fact be compatible with distinctiveness accounts (e.g., Ozubko et al., 2020; see also, Jamieson et al., 2016): Evidence suggests that the larger within-subject production effect results from both a benefit to aloud items and a *cost* to silent items (e.g., Bodner & Taikh, 2012; Bodner et al., 2014; Forrin et al., 2012; Ozubko et al., 2020). For example, Bodner et al. (2014) compared sensitivity scores (*d'*) for items that were studied in either a pure or mixed list. Sensitivity for items produced at study was shown to be similar across designs, whereas performance for silent items was significantly worse for mixed designs. This finding was replicated and extended by Ozubko et al. (2020), who found that adding encoding conditions to the study phase further hindered performance for silent items in mixed-list designs (see also, Forrin et al., 2012). Heuristic-based distinctiveness accounts dictate that in recognition

paradigms, participants search for a record of having produced a given item (i.e., distinctive information about encoding) to determine whether the item had previously been studied (see also, Dodson & Schacter, 2001). Ozubko et al. (2020) suggested that when participants fail to retrieve such a record – as would be expected for unproduced items – they are more likely to decide that an item was not studied, leading to a decrement in performance for silent items. When additional modalities are added at study, participants will also search for distinctive information relating to these other encoding conditions; if these checks also fail, participants' confidence in having studied silent items will decrease even further. Thus, a smaller between-subject production effect is congruent with some interpretations of distinctiveness: Because there is no cost to silent items in between-subject designs, the benefit for aloud items is larger relative to within-subject designs.

On the other hand, the existence of a smaller between-subject production effect also suggests that the benefit may not be explained by distinctiveness alone (Bodner et al., 2016; Fawcett & Ozubko, 2016). Key evidence for this assertion has been obtained from studies that have adapted production paradigms to test a *dual process* model (e.g., Fawcett & Ozubko, 2016). Generally, dual process models contend that performance on tests of memory is driven by two distinct processes: recollection and familiarity. Recollection is believed to involve a specific and vivid re-experiencing of the initial encoding episode, whereas familiarity is construed as more generalized and indirect feelings of having encountered the item before (for a review, see Yonelinas, 2002). Previous efforts that have tested the influence of production on each process by adding remember/know (i.e., recollect/familiar) judgements to the test phase in production paradigms have shown that production benefits both recollective and familiarity processes equivalently in within-subject designs (e.g., Ozubko et al., 2012b, 2014).

However, experiments by Fawcett and Ozubko (2016) demonstrated a pattern of influence that differed across designs such that the within-subject effect was driven by both recollection and familiarity, whereas the between-subject effect was driven by familiarity alone. Critically, the authors suggested that distinctiveness in production is a recollective process, given that participants are thought to consciously check their memories for specific details about the encoding episode (e.g., Dodson & Schacter, 2001; Forrin et al., 2012; MacLeod et al., 2010). Accordingly, Fawcett and Ozubko (2016) hypothesized that the larger within-subject production effect results from both distinctive recollective processes and another yet-to-be identified mechanism that drives the benefit in familiarity; on the other hand, the between-subject effect must be driven predominantly by the latter process alone, leading to a smaller benefit relative to the within-subject effect.[4] Although these findings suggest that distinctiveness does play a role in the within-subject benefit, they also imply that theoretical frameworks based solely upon encoding distinctiveness cannot account for the between-subject effect.

Such a hypothesis has been supported further by experiments that have tested source memory in between-subject production paradigms (e.g., Bodner et al., 2020). Production has been shown to benefit source memory in within-subject designs (e.g., Ozubko et al., 2012b; 2014), which is congruent with distinctiveness-based frameworks: Participants are thought to leverage distinctive information about the encoding episode (i.e., recollective processes) to guide memory for items, which should therefore translate to improved memory for details about the encoding episode (i.e., source). If the within- and between-subject production effects rely on the same type of distinctive processing at encoding, the benefit in source monitoring should persist

[4] Recent observations of a between-subject production effect in tests of recall (e.g., Cyr et al., 2021; Gionet et al., 2022; Saint-Aubin et al., 2021) – wherein test performance should depend predominantly upon recollective processes – suggest that the recollective component of the production effect may also involve processes beyond encoding distinctiveness.

regardless of experimental design. However, if the between-subject production effect relies instead on processing that is not recollective in nature – as suggested by Fawcett and Ozubko (2016) – production would not be expected to enhance source memory in these designs. A series of experiments by Bodner et al. (2020) confirmed the latter prediction, as production – when manipulated between-subject – did not enhance memory for item location, font size, or study modality.[5,6]

Another recent theoretical challenge to distinctiveness-based frameworks is the finding that the production effect can sometimes persist even when participants cannot leverage the production trace to guide discrimination between items (Fawcett et al., 2022). Fawcett et al. (2022) utilized a production paradigm that required two-alternative forced choice judgements at test between target-lure pairs that were either homophones (e.g., *band-banned*, *piece-peace*, etc.) or unrelated. According to typical distinctiveness accounts (e.g., Forrin et al., 2012; MacLeod et al., 2010), the production effect should be eliminated for participants tested using homophone lures: The distinctive auditory and motoric representations that production is thought to append to the memory trace would be ineffective in guiding discrimination between homophone pairs because either word in the pair would encode identical sensorimotor representations. Contrary to these predictions, Fawcett et al. (2022) found that the production effect persisted for participants tested with homophone lures and was similar in size to that observed for unrelated lures. Although previous studies that have experimentally manipulated the utility of the production trace have successfully eliminated the benefit altogether (e.g., MacLeod et al., 2010; Ozubko &

[5] In a subsequent experiment, Bodner et al. (2020) observed a production effect for judgements of study modality when manipulations of modality were made particularly salient to participants. However, some theorists have argued that judgements of study modality constitute *reality monitoring* (see Johnson et al., 1993 for a detailed discussion) and elicit different types of cognitive processing relative to external (i.e., perceptual) source monitoring tasks (see, e.g., Riefer et al., 2007). Thus, it is unclear whether the production effect in reality monitoring (see also, Ozubko et al., 2012b, 2014) is directly comparable to the external source monitoring tests in Bodner et al. (2020).

[6] In contrast to the above, production *does* seem to enhance memory for details beyond the item itself in within-subject designs (e.g., Hourihan & Churchill, 2020). However, some recent work suggests that these benefits might be driven by processes related to response generation rather than solely production (Whitridge et al., submitted).

MacLeod, 2010), the findings of Fawcett et al. (2022) indicate that the production effect can remain intact even when key elements of distinctiveness are obviated. Considered in aggregate with other theoretical challenges to the distinctiveness account (e.g., Bodner et al., 2020; Fawcett & Ozubko, 2016), these results strongly suggest that distinctiveness alone cannot fully explain the production effect. In the section that follows, I describe alternative theoretical explanations of the production effect and empirical findings that have supported – or failed to support – these accounts.

1.2 Alternative Accounts of the Production Effect

If distinctiveness alone cannot fully account for the benefit of production, what other cognitive mechanisms might drive the effect? One possibility is that produced items are simply better encoded relative to unproduced items, a hypothesis known as the *strength account*: If producing words lead to the formation of stronger memory traces, produced items should be more easily retrieved relative to unproduced items (MacLeod et al., 2010; Ozubko & MacLeod, 2010; Bodner & Taikh, 2012). Initially, this hypothesis was almost universally rejected by theorists because of failures to observe a significant between-subject production effect; strength accounts would predict production to benefit memory regardless of experimental design (MacLeod et al., 2010). Evidence for the between-subject effect, then, has renewed the viability of strength accounts (but see Jamieson et al., 2016), and a number of investigations have since provided evidence compatible with a role of encoding strength in facilitating production effects (e.g., Bodner et al., 2016, 2020; Fawcett & Ozubko, 2016; Icht et al., 2016; Mama & Icht, 2018; Taitelbaum-Swead et al., 2017).

Some evidence for a strength account has been derived from findings that manipulating statistical distinctiveness does not impact the magnitude of the production effect (Bodner et al.,

2016; *cf*. Icht et al., 2014; Zhou & MacLeod, 2021). In this context, statistical distinctiveness refers to the relative frequency with which a given type of item is processed (e.g., the percentage of aloud items in a mixed-list production paradigm) and can be distinguished from encoding distinctiveness, which refers to distinctive processing modalities employed at encoding (Huff et al., 2015; Gretz & Huff, 2020). Although distinctiveness accounts of the production effect typically refer – explicitly or otherwise – to encoding distinctiveness, the mixed-list advantage implies that statistical distinctiveness is also a factor (MacLeod et al., 2010; see also, Hopkins & Edwards, 1972). In a test of this account, experiments by Bodner et al. (2016) showed (1) that the size of the production effect did not differ across mixed and pure lists (but see Bodner et al., 2014), and (2) was consistent regardless of whether aloud items occurred 20% or 80% of the time in a mixed list. Distinctiveness-based accounts would predict the relative frequency of items to modulate the size of the production effect: When aloud items occur less frequently in a mixed list, they should be more distinctive relative to the backdrop of unproduced items and thereby better remembered. Accordingly, the findings of Bodner et al. (2016) are congruent with a strength account, which would predict a consistent production effect across differing degrees of statistical distinctiveness.

Further support for a strength account comes from findings that the production effect is augmented when production is more effortful (e.g., Icht et al., 2016; Mama & Icht, 2018; Taitelbaum-Swead et al., 2017). For example, Mama and Icht (2018) employed a paradigm wherein participants delayed their production of items until after the items had disappeared. Because the words were not visible at the time of production, participants had to retrieve items from working memory as a prerequisite to producing the items, increasing the amount of effort required to perform the task. Critically, Mama and Icht (2018) found that delaying production for

either one or three seconds resulted in a significantly larger benefit relative to immediate production. This finding cannot easily be accommodated by distinctiveness-based accounts: Both immediate and delayed production recruit the same distinctive encoding processes at study, and no additional sensorimotor processing that would be expected to give rise to a larger benefit occurs for delayed production. Moreover, some of the experiments conducted by Mama and Icht (2018) included only conditions wherein items were produced (rather than produced or read silently), eliminating any benefit that the relative distinctiveness of reading items aloud might afford.

Similarly, Icht et al. (2019) and Taitelbaum-Swead et al. (2017) respectively demonstrated that the production effect is larger in dysarthric and hearing-impaired populations relative to healthy controls. Because these clinical populations are characterized by difficulties with speech and hearing, respectively, production-related sensorimotor processing would be expected to require more effort relative to healthy populations. Once again, distinctiveness accounts cannot explain the effect's increase in magnitude under these circumstances, given that dysarthric and hearing-impaired participants would not be expected to recruit additional distinctive processing at encoding. However, the findings of Mama and Icht (2018; Icht et al., 2019; Taitelbaum-Swead et al., 2017) fit neatly within a strength account: When participants expend additional effort at encoding, the resulting memory trace should benefit from stronger encoding and be more easily retrieved, leading to a larger benefit.

Evidence for a strength account discussed thus far does not necessarily address the important questions of *why* produced items might be more strongly encoded. One candidate mechanism that has been identified in the literature is that the production effect might be driven by differential engagement with produced and unproduced items (e.g., Bailey et al., 2021;

Fawcett & Ozubko, 2016; Fawcett et al., 2022; MacDonald & MacLeod, 1998; Mama et al., 2018; Varao Sousa et al., 2013; Willoughby et al., 2019). According to this framework, participants attend better to produced items at study; allocation of additional cognitive resources to these items leads to better-encoded memory traces that can be more easily retrieved at test. Anecdotally, participants report paying more attention during study trials wherein items are produced (Fawcett & Ozubko, 2016) and additionally, the majority of evidence in support of a strength account is compatible with an attentional framework (e.g., Icht et al., 2019; Mama & Icht, 2018; Taitelbaum-Swead et al., 2017). However, the aforementioned findings do not necessarily implicate attention as the proximate cognitive mechanism underlying strength.

Other studies have provided more specific evidence for a role of attention in facilitating the production effect. For example, Varao Sousa et al. (2013) found that production appears to reduce distractibility. This study used a reading comprehension paradigm (see also, Ozubko et al., 2012a) with three conditions: Participants either read passages of text silently, aloud, or had the passages read to them. After studying the passages, participants self-reported instances of mind wandering that occurred during study. Importantly, the authors found that mind wandering was significantly lower when passages were read aloud relative to reading silently or passive listening. These results are incongruent with a distinctiveness account, given that there is no reason to expect that distinctive encodings at study would decrease distractibility. Instead, Varao Sousa et al. (2013) strongly supports an attentional framework: Participants allocate more attention to items during production and accordingly, fewer attentional resources are available to attend to distractions.

Further evidence for an attentional hypothesis comes from findings that the production effect is reduced or eliminated when decrements to attention are present (e.g., Mama & Icht,

2019; Mama et al., 2018). For example, Mama and Icht (2019) compared the production effect in healthy controls to that observed in participants with attention-deficit hyperactivity disorder (ADHD), a clinical population that typically exhibits severe impairments to executive function and sustained attention (Woods et al., 2002). The authors found that while participants with ADHD did exhibit a production effect, the benefit for aloud items was significantly smaller than that of the healthy controls. This is congruent with an attentional hypothesis insofar as deficits to attention appear to hinder the production effect, although this finding alone does not rule out alternative mechanisms that might occur due to other cognitive deficits associated with ADHD (e.g., working memory impairments). However, Mama and Icht (2019) also established that administration of methylphenidate – a drug commonly prescribed to treat ADHD – increased the size of the production effect in participants with ADHD such that performance for produced items was nearly equivalent to healthy controls. Critically, methylphenidate is thought to improve symptoms of ADHD because it normalizes attentional processing in neural networks that are typically disrupted for individuals with the disorder (Shafritz et al., 2004). It appears, then, that abnormal attentional processing disrupts the production effect, but the benefit returns to near-normal levels when this processing is normalized.

Similar findings have been obtained by experimentally manipulating attention in non-clinical populations using background noise (Mama et al., 2018). Task-irrelevant background noise impacts performance on a variety of cognitive tasks and this impact varies depending on how distracting the noise is (see Banbury et al., 2001 for a review). Steady-state, consistent background noise can typically be ignored and has little impact on task performance, whereas fluctuating noise disrupts attention and impairs performance. To test an attentional hypothesis of the production effect, Mama et al. (2018) had participants complete a typical production task

while either steady-state or fluctuating background noise was present. The authors found that the production effect was eliminated altogether when participants were exposed to fluctuating noise. Importantly, however, the production effect was robust in the steady-state noise condition, suggesting that the decrement observed for fluctuating noise did not occur simply because background noise interfered with distinctive auditory processing. Mama et al. (2018) explained these findings with reference to an attentional account: Fluctuating noise interfered with the increase in attention that is typically allocated to produced items, whereas steady-state noise was easily ignored and therefore left the increase in attentional processing intact. This pattern of results is akin to that observed by Mama and Icht (2019) and provides direct evidence supporting a role of attention in facilitating the production effect.

An attentional account of the production effect has also been investigated using neurocognitive production paradigms, which have provided mixed evidence supportive of a role for attention (e.g., Bailey et al., 2021; Hassall et al., 2016; Zhang et al., 2023). The first such investigation was conducted by Hassall et al. (2016), who used electroencephalography (EEG) to record neural activity during study trials in a three-condition production paradigm (i.e., read aloud, read silently, sing aloud). Activity was measured at 300-500 ms following presentation of the instruction to produce or silently read an item (i.e., during the preparatory phase of the productive act). In addition to a typical behavioral production effect (i.e., better recognition performance for produced versus unproduced items), the authors observed a pattern of psychophysiological results that mirrored the behavioral findings: Relative to trials for which items were read silently, production trials were associated with increases to the amplitude of the P300b (P3b) component of the event-related brain potential. Some studies have interpreted such a pattern in P3b amplitude as reflective of distinctive encoding. For example, Otten and Donchin

(2000) and Karis et al. (1984) demonstrated increases to the amplitude of the P3b for words presented in large (relative to small) font, a manipulation thought to render items more distinctive. Similarly, P3b amplitude has been shown to scale proportional to the statistical distinctiveness of a given type of item presented in a set (e.g., Donchin, 1981). Accordingly, Hassall et al. (2016) interpreted their findings as supportive of a distinctiveness hypothesis: Producing words elicits distinctive processing and thereby produces a larger P3b.

However, other evidence suggests that the P3b can reflect increases in attentional allocation at encoding (see Kok, 2001, and Polich & Kok, 1995, for reviews). For example, the amplitude of the P3b can be increased by (1) manipulating task difficulty – and thereby the degree of attentional processing required for task performance – by introducing a second, distracting task (e.g., Isreal et al., 1980); and (2) increasing the salience of task-relevant (but not irrelevant) stimuli in divided attention paradigms (e.g., Kramer et al., 1983). Thus, the patterns of results reported by Hassall et al. (2016) is also compatible with an attentional account of the production effect. This hypothesis is favored by Zhang et al. (2023), who replicated the psychophysiological results of Hassall et al. (2016) using a similar paradigm. Zhang et al. (2023) were additionally able to rule out the notion that increases to the P3b might have resulted simply because participants had to respond vocally to aloud trials. This was accomplished through the inclusion of an aloud control condition, wherein participants responded to all items by producing the word “check” aloud; for these trials, there was neither a behavioral nor psychophysiological production effect (see also, MacLeod et al., 2010; Richler et al., 2013). Accordingly, increases to the P3b during the preparatory phase of production appear to reliably indicate a production-related change in processing at encoding, but the matter of whether this change in processing reflects distinctiveness or attentional increases cannot yet be resolved (Zhang et al., 2023).

However, it should be noted that the P3b is generally interpreted as an index of attentional allocation rather than distinctive encoding (Kok, 2001; Polich & Kok, 1995).

In addition to EEG paradigms, an investigation of the production effect using functional magnetic resonance imaging (fMRI) conducted by Bailey et al. (2021) showed patterns of neural activity compatible with an attentional hypothesis. In this study, participants completed a typical production task while situated within an MRI machine, and functional neural scans were obtained prior to and during the study and test phases; Bailey et al. (2021) also included an aloud control condition similar to that described in Zhang et al. (2023) to disentangle changes in activity that might occur when a response of any kind is made from changes specific to production of unique items. The authors found that producing items at study was associated with increased neural activation in areas associated with auditory and articulatory processing (e.g., auditory cortex, premotor cortex), which is congruent with suggestions that participants allocate additional attention to sensorimotor processes during production. However, Bailey et al. (2021) note that these findings are also consistent with distinctiveness-based hypotheses: Increased activation in regions relevant to sensorimotor processing might also reflect distinctive processing. Much like evidence derived from EEG studies of production, then, fMRI cannot yet adjudicate between theoretical accounts of the effect.

Finally, one recent investigation used pupillometry as an indirect index of attentional allocation to explore the role that attention might play in facilitating the production effect (Willoughby, 2019). Cognitive pupillometric studies use eye-tracking devices to measure pupil dilation while participants perform cognitive tasks; evidence generally suggests that pupil dilation increases as participants expend more cognitive effort during task performance (see Beatty, 1982 for a review). If production increases attentional engagement, then, pupil dilation

would be expected to increase during the productive act. To test this hypothesis, Willoughby et al. (2019) had participants complete a typical production task (including an aloud control condition) while being monitored by an eye tracker. In line with an attentional account, the authors observed a significant increase in pupil dilation for aloud relative to silent trials. However, this benefit occurred in both the aloud and aloud control conditions; as such, increases in pupil dilation may have occurred simply due to the increased cognitive effort associated with responding to trials vocally. Because no behavioral production effect was observed in the aloud control condition, this increase in cognitive effort cannot be definitively linked to the benefit. However, Willoughby et al. (2019) also observed a decrease in pupil size during silent trials such that dilation dropped below baseline measures late in the trials. This finding may suggest that participants engage in less effortful processing during silent trials, which could potentially explain the cost to silent items observed in mixed list designs (e.g., Bodner et al., 2014; Ozubko et al., 2020) and favors a role for attentional processes in facilitating the production effect. Given the inconsistent findings for aloud trials, however, pupillometric evidence for an attentional account is mixed at present.

Considered in aggregate, the behavioral and neurocognitive evidence discussed thus far provides a solid basis for an attentional account of the production effect: Even in instances where an attentional account was not necessarily directly supported, the account certainly cannot be ruled out (e.g., Bailey et al., 2021; Willoughby et al., 2019). However, some discrepant evidence remains. For example, two experiments reported in MacDonald and MacLeod (1998) used a production paradigm wherein participants completed both direct (i.e., explicit) and indirect (i.e.,

implicit) tests of memory, with no benefit of production being observed for the latter.[7] For the implicit test, MacDonald and MacLeod (1998) used a rapid reading paradigm (MacLeod, 1996). In this task, words are presently rapidly at test and are intermixed with new distractor words (i.e., foils). Participants are instructed to read the words as quickly as possible at test, with the response latencies for previously studied items serving as a measure of implicit memory: Words that were previously learned should be primed in memory and should thereby result in lower response latencies relative to new items. Further, response latencies should decrease if some words were learned better than others. As such, distinctiveness accounts would not predict production to benefit performance on rapid reading tests: Because this paradigm relies on near-automatic processing, application of a distinctiveness heuristic – which is explicit and time-consuming – would serve no purpose (MacLeod et al., 2010). On the other hand, participants should respond faster to aloud items if they allocate more attention to these items at study (e.g., through stronger encoding and better priming).

Contrary to predictions made by attentional accounts, MacDonald and MacLeod (1998) observed similar performance on the indirect memory test across the silent and aloud conditions, despite observing a typical production effect on the direct memory test. However, differences in attentional allocation at study do not consistently produce differences in performance on indirect tests of memory (e.g., Jacoby et al., 1989; Kellogg et al., 1996). To account for this inconsistency, Mulligan and Hartman (1996) suggested that manipulating attention at study impacts indirect test performance only when the test relies on conceptual – rather than perceptual – processing. Because MacDonald and MacLeod's (1998) rapid reading test necessitated only

[7] MacDonald and MacLeod's (1998) first experiment had participants read some words aloud and respond to others with "pass," rather than silent reading. However, this manipulation typically results in a production effect (i.e., better performance for aloud relative to aloud control trials) despite the need for an aloud response on "pass" trials (e.g., Bailey et al., 2021; MacLeod et al., 2010; Willoughby, 2019; Zhang et al., 2023).

that participants visually process the word and quickly repeat it, this task is predominantly perceptual in nature (for detailed discussion of conceptual and perceptual processing in cognitive tasks, see Jacoby, 1983 and Mulligan, 2011). Accordingly, whether production-related differences in attentional allocation would actually be reflected in rapid reading performance is dubious. Nonetheless, the findings of MacDonald and MacLeod (1998) have been accepted as evidence against an attentional account (see MacLeod et al., 2010).

Thus far, the evidence reviewed for various theoretical frameworks of the production effect is far from conclusive. Some studies support a role of distinctiveness (e.g., Forrin et al., 2012; MacLeod et al., 2010; Ozubko & MacLeod, 2012; Richler et al., 2013), whereas others favor alternative explanations (e.g., Fawcett & Ozubko, 2016; Fawcett et al., 2023; Mama & Icht, 2019; Mama et al., 2018) or are compatible with multiple frameworks (e.g., Bailey et al., 2021; Hassall et al., 2016; Zhang et al., 2023). An important point that has been raised in the literature is that these accounts are not necessarily mutually exclusive (e.g., Bodner et al., 2020; Fawcett, 2013; Fawcett & Ozubko, 2016; Fawcett et al., 2023; Jamieson et al., 2016). For example, items could be more strongly encoded because of distinctive processing, or distinctive encoding could result from increased attentional allocation to sensorimotor processing. Jamieson et al. (2016) noted the difficulties in computationally modelling a strength account of the production effect that differs qualitatively from a feature-based distinctiveness account: If one accepts that both distinctiveness and strength accounts benefit memory by appending additional features to the memory trace – features which are either more numerous or better encoded, respectively – then the mechanisms will prove difficult to disentangle experimentally.

Nonetheless, the majority of empirical evidence presently available favors a distinctiveness account over other explanations. To summarize my earlier discussion, support for

the distinctiveness framework derives largely from three key points: first, that the production effect can be eliminated by obviating the diagnostic value of the production trace (e.g., Ozubko & MacLeod, 2010); second, that the production effect does not occur if participants employ a non-distinct response at study (e.g., MacLeod et al., 2010; Richler et al., 2013); and finally, that the production effect is less pronounced in pure-list designs (e.g., Fawcett, 2013; Fawcett et al., 2023). Conversely, several key findings within the production literature instead argue against a predominant role for encoding distinctiveness, including evidence that the production effect can persist even when the production trace is rendered non-distinctive (e.g., Fawcett et al., 2022), and that the recollective component of the benefit is absent in pure list designs (e.g., Fawcett & Ozubko, 2016; see also, Bodner et al., 2020). Finally, although alternative theoretical perspectives of the production effect have not received a great deal of exclusive support, many investigations have produced evidence compatible with these frameworks: Studies have shown that the production effect can be reduced or eliminated by manipulating attention at encoding (e.g., Mama & Icht, 2019; Mama et al., 2018) and that production appears to elicit increases in preparatory processing (e.g., Hassall et al., 2016; Willoughby et al., 2019) and reductions in mind wandering (Varao-Sousa et al., 2013).

1.3 Scaling Distinctiveness in the Production Effect

The central purpose of the present thesis is to investigate one important prediction of the distinctiveness account, which I term the *sensorimotor scaling hypothesis*. Most models of distinctiveness contend that the presence of additional distinctive sensorimotor features associated with the production trace drives the production effect (e.g., MacLeod et al., 2010; see also, Kappel et al., 1973). Thus, a key corollary of this framework – which was first proposed by Forrin et al. (2012) and later adopted elsewhere (e.g., Fawcett et al., 2012; Hassall et al., 2016;

Jamieson et al., 2016; Kelly et al., 2022; Mama & Icht, 2016; Quinlan & Taylor, 2013, 2019) – is that the magnitude of the production effect should scale proportionate to the number of distinctive features encoded at study. Reading words aloud would be expected to enlist three distinct sensorimotor processes at study: visual processing of the item during reading, motoric processing that occurs when participants move their mouths to produce the item, and auditory processing occurring when participants hear themselves say a word. Silent reading, on the other hand, recruits only visual processing (see also, Forrin & MacLeod, 2018; Mama & Icht, 2016). According to the sensorimotor scaling hypothesis, then, the production effect should decrease in magnitude for tasks that reduce the number of distinctive features encoded at study.

A number of investigations have provided empirical support for such a prediction (e.g., Conway & Gathercole, 1987; Forrin et al., 2012; Mama & Icht, 2016; Taitelbaum-Swead et al., 2018). For example, Conway and Gathercole (1987) tested participants in a three-condition production paradigm wherein words at study were either read silently, mouthed or read aloud. According to the sensorimotor scaling hypothesis, mouthing would be expected to recruit only visual and motoric processing – rather than visual, motoric, and auditory processing – at study; as such, the production trace resulting from this modality should possess fewer distinctive features to guide retrieval at test relative to those produced by reading aloud. Congruent with the model proposed by Forrin et al. (2012), Conway and Gathercole (1987) observed production benefits for both reading aloud and mouthing but found the benefit for the former to be significantly larger than that for the latter. Later investigations by Forrin et al. (2012) replicated this finding and examined whether the result would extend to the production effect for writing; like mouthing, writing would be expected to recruit only visual and motoric processing at study, which should result in fewer distinctive features being appended to the production trace and

thereby a smaller production effect. Using a similar three-condition production paradigm (i.e., read silently, read aloud, write), Forrin et al. (2012) found that the production effect for writing was significantly smaller than that for reading aloud. It appears, then, that the size of the production effect is moderated by the number of distinctive features encoded by the production modality employed at study.

Subsequent extensions of these findings have demonstrated that both input (i.e., presentation) and output (i.e., production) modality at study can impact the relative superiority of different modes of production (e.g., Mama & Icht, 2016; Taitelbaum-Swead et al., 2018). For example, Mama and Icht (2016) manipulated the modality through which items were presented (i.e., visually or auditorily) in a production paradigm that included both reading aloud and writing as output modalities. Although writing would be expected to recruit only visual and motoric processing in a typical (i.e., visual) paradigm (Forrin et al., 2012), auditory processing should also be recruited when words are presented auditorily. On the other hand, presenting words in this manner would eliminate the distinctive visual features that are usually encoded for words that are visually presented and vocally produced. For auditory presentation, then, vocal production should append a smaller number of distinctive features to the production trace (i.e., auditory and motoric features) relative to written production (i.e., auditory, motoric, and visual features). Accordingly, a scaling model of distinctiveness would predict a larger production effect for reading aloud when words are presented visually, but the production effect for writing should be superior when presentation is auditory. The results of Mama and Icht (2016) fully supported this prediction: The vocal production effect was significantly larger than that observed for writing when presentation was visual, but this pattern was reversed for auditory presentation.

Similar results were obtained by Taitelbaum-Swead et al. (2018), who investigated the production effect in hearing impaired users of sign language. At study, participants studied words that were presented either visually (i.e., written) or manually (i.e., using sign language) and produced the words via signing; the authors found that the production effect was significantly smaller for words presented manually. For both visual and manual presentation, words would be expected to benefit from visual processing necessitated by the input modality and motoric processing necessitated by the output modality. However, evidence suggests that signing – particularly for fluent users of the language – involves additional distinct forms of processing that are not solely visual nor motoric (e.g., Wilson & Emmory, 1997). Interpreting these results in light of a sensorimotor scaling hypothesis, Taitelbaum-Swead et al. (2018) concluded that manually presenting words at study reduced the advantage for produced items because the sensory processing specific to signing occurred on all trials; for visually presented words, on the other hand, these additional processes could still be leveraged to better discriminate between produced and unproduced items. This pattern mirrors that observed by Mama and Icht (2016): Manipulating the input modality such that certain distinctive sensorimotor features are encoded for both produced and unproduced items can moderate the production effect. Accordingly, evidence derived from manipulation of both output and input modalities converges in favor of the sensorimotor scaling hypothesis and, by extension, the distinctiveness account: As the number of distinctive features that can be leveraged to guide retrieval at test decreases, so too does the magnitude of the production effect.

The sensorimotor scaling hypothesis would also predict the inverse of this pattern to occur: Adding distinctive features to the production trace by recruiting additional forms of sensory processing at study should increase the magnitude of the production effect (Fawcett et

al., 2012; Forrin et al., 2012). In an initial attempt to test this prediction, Fawcett et al. (2012) compared memory for produced and unproduced stimuli that were either words or pictures. Typically, participants exhibit a mnemonic advantage for pictures relative to words, a finding termed the *picture superiority effect* (Paivio, 1991). Much like the production effect, the advantage for pictorial stimuli has been explained by a distinctiveness heuristic, wherein participants leverage distinctive visual features associated with the memory trace to guide retrieval at test (Lloyd & Miller, 2011; Schacter et al., 1999). Accordingly, producing the names of pictures at study should encode additional distinctive features, resulting in a larger production effect than that observed for words. In line with a sensorimotor scaling hypothesis, Fawcett et al. (2012) observed an interaction between production and stimulus type across three experiments such that the production effect was consistently larger for pictures relative to words.

However, subsequent attempts to replicate this pattern of results have produced inconsistent evidence. Experiments conducted by Zormpa et al. (2019) compared performance for words and pictures that were either produced or unproduced. In their first experiment, Zormpa et al. (2019) successfully replicated the interaction observed by Fawcett et al. (2012). However, the authors speculated that when pictorial stimuli are used at study, the benefit of production is confounded with that of response generation (e.g., Slamecka & Graf, 1978; see McCurdy et al., 2020 for a meta-analytic review of generation effects in memory): When pictures are presented without labels – as in Fawcett et al. (2012) – participants must generate the name of the picture in order to produce it. To disentangle these processes, Zormpa et al. (2019) manipulated the presence of labels with the pictorial stimuli in a subsequent experiment. The authors found that when the confounding influence of response generation was eliminated via the inclusion of labels, the critical interaction between production and picture superiority was not

observed. Furthermore, a recent study by MacLeod et al. (2022) found no such interaction even when unlabelled pictures were used at study. Accordingly, the support that Fawcett et al. (2012) provides for the sensorimotor scaling hypothesis is not necessarily reliable and may instead be attributable to artifacts related to response generation.

Recently, Wakeham-Lewis et al. (2022) attempted to increase the distinctiveness of the production trace by having participants produce items using either unusual voices (e.g., character voices) or their own voice. The authors argued that because neuroimaging evidence suggests that voluntary modulation of one's own voice recruits processing in neural regions that differ relative to normal speech (McGettigan et al., 2013), speaking in an unusual voice represents a distinct form of processing. Compared to normal speech, then, the memory trace resulting from producing items in an unusual voice should benefit from additional sensorimotor features; the production effect for this modality would therefore be expected to increase in magnitude relative to production using one's own voice. Contrary to these predictions, however, Wakeham-Lewis et al. (2022) actually observed no benefit of production whatsoever when unusual voices were employed at study. However, whether this finding speaks directly against a sensorimotor scaling hypothesis is unclear at present. The authors suggested that the utility of the production trace might depend to some extent on reinstating study context (i.e., of having produced the item) at test; if producing items in an unusual voice interferes with this reinstatement, the benefit of production might be eliminated. Nonetheless, the results of Wakeham-Lewis et al. (2022) suggest that employing modalities expected to invoke additional distinctive processes at study will not necessarily translate to a larger production effect.

Accordingly, the sensorimotor scaling hypothesis has largely been validated only insofar as the magnitude of the production effect can be reduced, whereas modalities that produce a

larger benefit remain scant. However, one notable exception to this pattern is the production effect for singing. Across multiple experiments, Quinlan and Taylor (2013) observed a benefit for singing that was significantly larger than that for reading words aloud; I refer to this hereafter as the *singing superiority effect*. The authors explained this finding with reference to the sensorimotor scaling hypothesis, speculating that singing appends additional sensorimotor features to the production trace in the form of processing related to pitch, tone, or rhythm.[8] This finding has since been accepted as evidence for the distinctiveness account (e.g., Forrin & MacLeod, 2018; Mama & Icht, 2016) by virtue of apparently validating the sensorimotor scaling hypothesis proposed by Forrin et al. (2012).

Further investigations by Quinlan and Taylor (2019) replicated the initial finding of a singing superiority effect and extended the result in attempts to rule out a number of alternative explanations beyond distinctiveness. First, the authors tested whether the additional benefit observed for singing might result from a bizarreness effect (see, e.g., Einstein et al., 1987) by recruiting a sample of experienced singers, for whom the authors reasoned singing words would not constitute an unusual task. The production effect for singing was found to be robust to this manipulation and remained significantly larger than the standard vocal production effect. In a subsequent experiment, Quinlan and Taylor (2019) determined that singing words at study is more time-consuming than reading aloud, potentially allowing participants additional time to better encode the words. However, the singing superiority effect was found to persist even when the processing time-related advantage was eliminated by having participants sing words quickly and read words aloud slowly. Finally, the authors investigated whether words sung at study might

[8] Importantly, however, features related to pitch, tone, and rhythm are not necessarily specific to singing and are also present for speech (e.g., Dolson, 1994; Xu, 2005). Thus, although the singing superiority effect has often been explained with reference to additional sensory processing related to pitch (e.g., Quinlan & Taylor, 2013, 2019), such a hypothesis may not actually provide a viable theoretical basis for the effect (see Section 5.3 of the present thesis for further discussion).

be better remembered due to strengthened encoding. To this end, Quinlan and Taylor (2019) manipulated production in a between-subject design; here, the authors did not observe a singing superiority effect, nor a production effect for either reading aloud or singing (but see Bodner et al., 2014; Fawcett, 2013; Fawcett & Ozubko, 2016; Fawcett et al., 2023). Having apparently ruled out all viable alternatives, Quinlan and Taylor (2019) reasoned that the relative superiority of singing over standard vocal production could only be explained by scaling distinctiveness.

However, not all investigations into the production effect for singing have observed a singing superiority effect. One EEG study of the production effect by Hassall et al. (2016) included singing as a condition. In this case, however, the authors observed a behavioral production effect of singing on recognition memory that was similar in size to the benefit for reading aloud. Furthermore, the psychophysiological production effect reflected in increases to P3b amplitude conformed to the same pattern. Because the P3b has variably been used as index of attentional allocation (e.g., Kramer et al., 1983) and distinctive encoding (e.g., Otten and Donchin, 2000), these results suggest that items sung at study are neither better attended to nor more distinctive than those read aloud. However, the authors proposed that these findings did not necessarily contradict a distinctive account, proposing instead two alternative explanations: (1) that participants might have failed to follow task instructions (e.g., by lazily singing the items), or (2) that the temporal delay associated with pre-cueing production – which was necessitated by the EEG paradigm in order to reduce data contamination – might have reduced the distinctiveness of the production trace.

Although the former explanation is plausible, participants in Hassall et al. (2016) were monitored by an experimenter throughout the study phase. Accordingly, it is unclear why the failure of participants to adequately follow task instructions would not have been noticed and

addressed prior to analysis of the data. Regarding a "temporal separation" explanation, the authors suggested that distinctive cues arise from processing that occurs during both the intention to perform the productive act and the act itself. The psychophysiological results of both Hassall et al. (2016) and Zhang et al. (2023) support the notion that participants engage in some form of preparatory processing prior to actually producing the item (see also, Willoughby et al., 2019). Further, these findings are compatible with the suggestion that this preparatory processing might reflect distinctive encoding (see Section 1.2 for further discussion). However, it is not clear why a small temporal separation (1800 – 2000 ms) between intention and action would reduce the distinctiveness of either component. Several studies have pre-cued production with similar temporal separations and observed typical production effects (e.g., Bailey et al., 2021; Ozubko et al., 2020; Zhang et al., 2023). Furthermore, Mama and Icht (2018) found that the magnitude of the production effect *increased* when the intention and act of production were separated by an even larger delay (i.e., 3000 ms). Finally, this apparent reduction in distinctiveness appeared to preferentially impact singing: The magnitude of the production effect for reading aloud that Hassall et al. (2016) – and other investigations that have pre-cued production at study (e.g., Bailey et al., 2021; Ozubko et al., 2020; Zhang et al., 2023) – observed was similar in size to those observed in typical studies (e.g., MacLeod et al., 2010). Given that the production effect for reading aloud is also thought to be driven by encoding distinctiveness, there is no obvious reason why a reduction in distinctiveness would not have also altered the magnitude of the effect for this modality. With these considerations in mind, the results of Hassall et al. (2016) appear to largely contradict those of Quinlan and Taylor (2013, 2019).

More recent efforts to replicate the singing superiority effect have also proven unsuccessful. Whitridge (2022) conducted two conceptual replications of Quinlan and Taylor's

(2013) paradigm wherein production was manipulated within-subject. Whitridge's (2022) first experiment demonstrated a pattern of results akin to Hassall et al. (2016): The author observed a production effect for both singing and reading aloud but provided evidence against a difference between the two critical conditions. In a subsequent experiment, Whitridge (2022) attempted to rule out hidden moderators related to study design as a potential explanation for this pattern of results. At test, Quinlan and Taylor (2013, 2019; Hassall et al., 2016) presented studied items using the same color assignments as the study phase, with foil items randomly intermixed between color assignments; this foil matching procedure allows separate false alarm rates for each condition to be recorded in order to permit a complete signal detection analysis of the data (see Fawcett et al., 2012), but is atypical for production paradigms. Whitridge (2022) reasoned that participants might have leveraged familiar color assignments as contextual cues to guide retrieval of condition-specific distinctive information. When the presence of color matching was manipulated between-subject, however, the author again observed evidence against a singing superiority effect across groups.

In an attempt to resolve this apparently discrepant pattern of results, Whitridge (2022) conducted a meta-analysis of all known studies to have evaluated the production effect for singing. The author found that the aggregate singing superiority effect was significant (albeit small) but also observed evidence of substantial heterogeneity amongst reported effects: The aggregate effect appeared to be driven primarily by underpowered studies that reported very large effect sizes. On the other hand, well-powered experiments reported effects that were typically small and sometimes non-significant; in some cases, larger studies actually reported negative effects (i.e., read aloud > sing). Based on this evidence, Whitridge et al. (2022) concluded that the singing superiority effect was unreliable and questioned whether there exists a

viable cognitive basis – pertaining to distinctiveness or otherwise – for the relative superiority of singing over reading aloud. In the following section, I review literature that has investigated the utility of singing as a mnemonic and discuss processing differences that might provide a theoretical basis for the singing superiority effect.

1.4 Theoretical Bases of Singing as Mnemonic

Although Quinlan and Taylor (2013) were the first to investigate singing in the context of the production effect, a great deal of earlier literature has explored the mnemonic benefit that song might afford (e.g., Anton, 1990; Gfeller, 1983; Prickett & Moore, 1991; Richards, 1969; Wallace, 1994). Within this area of research, one oft-cited finding is Wallace's (1994) observation of better recall for words set to a melody relative to spoken words. In this study, Wallace (1994) had participants listen to an entire song that was either sung or spoken and then attempt to recall the song's lyrics; this procedure was repeated five times, although recall performance was reported only for the first, second and fifth trials. At all time points, recall was superior for the sung condition relative to the spoken condition. Furthermore, Wallace (1994) found that this benefit persisted even after a substantial delay during which filler tasks were completed (see also, Good et al., 2015). The author hypothesized that this advantage may have occurred because participants used melodic information as a retrieval cue at test. This explanation is analogous to a distinctiveness account, in that Wallace (1994) suggests a record of distinctive sensory processing that occurred at encoding is accessed at test and leveraged to guide retrieval of items.

However, subsequent investigations have largely failed to further support these findings (e.g., Kilgour et al., 2000; Rainey & Larsen, 2002; but see Salcedo, 2010). In an initial experiment, Kilgour et al. (2000) successfully replicated the advantage for listening to words that

were sung relative to spoken words. However, Kilgour et al. determined that the rate of presentation for words in the sung condition was much slower than that in the spoken condition, potentially allowing participants additional time to process and rehearse the information (see also, Quinlan & Taylor, 2019). In a second experiment that increased the speed of sung words to match that of the spoken words, the advantage for the former was eliminated. Furthermore, Rainey and Larsen (2002) tested memory in a similar paradigm, wherein participants heard novel words that were either spoken or sung to familiar melodies. Contrary to Wallace (1994), performance on an immediate serial recall task was no better for sung words in either of the two experiments conducted.[9] It appears, then, that processing melodies at study may not inherently provide distinctive information that can be leveraged at test. Thus, the extent to which studies like Wallace (1994) can provide a theoretical basis for singing superiority effects in production paradigms is tenuous.

Nonetheless, these studies differ critically from Quinlan and Taylor (2013, 2019; Hassall et al., 2016) insofar as participants passively listened to items rather than actively producing them. Although evidence suggests that input modality can be important to the production effect (e.g., Mama & Icht, 2016; Taitelbaum-Swead et al., 2018), the benefit is driven predominantly by self-producing items (MacLeod, 2011). Although few studies beyond those already discussed have tested memory directly for information sung by participants at study, some investigations of second-language vocabulary acquisition might have particular relevance to the production effect (e.g., Baills et al., 2021; Ludke et al., 2014).

[9] Rainey and Larsen (2002) observed an advantage in serial recall for words presented in the sung condition after a delay of one week, suggesting that the presence of melody can improve memory to some degree. However, the mechanism through which this benefit occurred was evidently not immediately available to participants at test; given that only one production study has included a delay of this length between study and test (Ozubko et al., 2012b), it is unlikely that the delayed benefit observed by Rainey and Larsen (2002) would transfer to typical production tasks.

For example, Ludke et al. (2014) tested participants' memory for English-Hungarian paired associate phrases that were presented auditorily in one of three conditions: sung, read aloud, or read aloud rhythmically. For each condition, participants were instructed to reproduce the stimuli in the same manner that they were initially presented. The authors found that performance on several different memory tests for both English and Hungarian phrases was better for those sung at study relative to either spoken condition, for which performance was equivalent (*cf*. Bailles et al., 2021).[10] Although the paradigm used differed substantially from typical production tasks, these findings are consistent with those reported by Quinlan and Taylor (2013, 2019); thus, Ludke et al. (2014) provides additional evidence in favor of a singing superiority effect. Furthermore, the pattern of results observed therein (i.e., sing > read aloud = read aloud rhythmically) raises an important implication, in that distinctive processing at study related to rhythmic information does not appear to benefit memory; if a reliable singing superiority effect exists, then, this advantage might instead be related to other types of processing.

Indeed, the specific nature of processing that occurs during singing is important to consider in establishing a theoretical basis for the singing superiority effect; Quinlan and Taylor (2013, 2019) suggested that singing appends additional distinctive features to the production trace, which implies that singing elicits fundamentally different processing relative to reading aloud. Generally, this assertion has been supported empirically. Neuroimaging studies suggest that human speech and song elicit increased neural activation in both the superior temporal gyrus (STG) and superior temporal sulcus (STS) regions of the brain (e.g., Özdemir et al., 2006;

[10] Interestingly, Bailles et al. (2021) observed no production effect in a foreign language acquisition paradigm for a passage that was sung at study relative to passive listening of the same passage. Although the relative superiority of singing over reading aloud has been inconsistent in the literature, a typical production effect for singing has not (Whitridge et al., 2022). Accordingly, it is likely that the findings of Bailles et al. (2021) can be attributed to substantial methodological differences (e.g., stimulus complexity).

Whitehead & Armony, 2018). While activation in the STS is greater for spoken language relative to singing, the latter elicits increased activation in the STG. This differential pattern of activation persists even when accounting for the pitch and rhythm of vocalization (e.g., Geiser et al., 2008; Özdemir et al., 2006), implying an inherent difference between how spoken and sung words are processed. Accordingly, increased activation in the STG provides a potential neural correlate for the singing superiority effect: If this activation reflects additional distinctive processing of elements unique to or more pronounced for singing (e.g., melody), the sensorimotor scaling hypothesis would indeed predict a singing superiority effect. With this in mind, however, it is important to consider that production modalities neurologically distinct from reading aloud do not inherently translate to a larger production effect (Wakeham-Lewis et al., 2022).

Considered in aggregate, singing-related mnemonic benefits and differences in neurological activation provide some degree of theoretical basis for the singing superiority effect. However, whether this basis can be realized reliably in production paradigms remains unclear (Whitridge, 2022). Given that the sensorimotor scaling model of distinctiveness proposed by Forrin et al. (2012) has been widely adopted into theoretical perspectives of the production effect (e.g., Fawcett et al., 2012; Forrin & MacLeod, 2018; Jamieson et al., 2016; Kelly et al., 2022), the reliability of the singing superiority effect is a critical issue. Although several studies suggest that the magnitude of the production effect decreases in proportion to the number of distinctive features encoded at study (e.g., Mama & Icht, 2016; Taitelbaum-Swead et al., 2018), the flipside of this prediction relies almost entirely on the finding of a singing superiority effect. A challenge to this pattern of results thereby poses a challenge to the distinctive account of the production effect: If the singing superiority effect is not reliable, the sensorimotor scaling hypothesis can only be verified unidirectionally. Despite the importance of this issue, it is difficult to draw

strong conclusions about the reliability of the singing superiority effect given the scarcity of relevant literature. In the investigations described below, I address gaps in previous conceptual replications of the production effect for singing (Whitridge, 2022) and attempt to further elucidate the phenomenon.

1.5 Current Experiments

The present investigations were designed to address several key issues regarding the production effect for singing. First, findings from the few studies that have included singing as a manipulation in production tasks have been inconsistent and often contradictory. To explain this discrepancy amongst results, Quinlan and Taylor (2019; Hassall et al., 2016) suggested that methodological differences might have obviated the singing superiority effect. Alternatively, several key experiments utilized very small sample sizes (e.g., < 24 participants; Quinlan & Taylor, 2013), which Whitridge (2022) speculated might have led to poor estimates of the effect's magnitude; these possibilities were explored in Experiment 1. Second, the failure of Quinlan and Taylor (2019) to observe a between-subject production effect for singing is inconsistent with a growing body of literature suggesting that the production effect is reliable in such designs (e.g., Bodner et al., 2014; Fawcett, 2013; Fawcett & Ozubko, 2016; Fawcett et al., 2023). Given that the magnitude of the between-subject production effect is smaller than that of the effect within-subject, I speculated that the results of Quinlan and Taylor (2019) may again have resulted from inadequate statistical power; Experiment 2 investigated this hypothesis. Finally, the overall disagreement in findings across studies of the production effect for singing was addressed in the meta-analysis.

Experiment 1 was designed as a replication and extension of Whitridge (2022). Although the aforementioned study provided evidence against artifacts related to one particular aspect of

study design (i.e., color matching), other aspects of the paradigm differed from that used by Quinlan and Taylor (2013, Experiment 1). Accordingly, the materials and procedure for Experiment 1 were a modified version of that used in Whitridge (2022, Experiment 2) such that the design of the study matched Quinlan and Taylor (2013) as closely as possible; additionally, the presence of foil matching at test was manipulated between-subjects. In Experiment 2, I conceptually replicated the paradigm employed in Experiment 4 of Quinlan and Taylor (2019): Production modality (sing, aloud, silent) was manipulated between-subjects, such that participants studied a pure list of items in a manner corresponding to the group to which they were assigned. Finally, the meta-analytic model incorporated data from all known studies to investigate the production effect for singing. Effect sizes obtained from these data were used to calculate an aggregate singing superiority effect, which could then be evaluated for evidence of heterogeneity.

Across Experiments 1 and 2, I expected a pattern of results consistent with Hassall et al. (2016) and Whitridge (2022). Relative to silent reading, a production effect should occur for both singing and reading aloud; however, I expected to observe no difference between the two critical conditions (i.e., sing = aloud > silent). Further, I expected this pattern of results to persist for judgements of both recollection and familiarity in Experiment 2. Finally, I predicted that the meta-analysis would support a small aggregate singing superiority effect, albeit with evidence of heterogeneity, as in Whitridge (2022). Although this may appear discrepant with my predictions for Experiments 1 and 2, I assumed that the large, supportive effects reported in Quinlan and Taylor (2013, 2019) would drive a credible benefit in aggregate even if new effects derived from the current experiments were unsupportive.

Chapter 2: Experiment 1

2.1 Overview

The central purpose of this experiment was to determine whether previous efforts to replicate the singing superiority effect failed due to methodological differences. Previously, Whitridge (2022) found evidence against a singing superiority effect regardless of whether items were foil matched or unmatched at test. However, a number of other differences existed between Quinlan and Taylor (2013, Experiment 1) and the replications conducted by Whitridge (2022). For example, Quinlan and Taylor (2013) included practice and familiarization phases prior to the study phase; it is possible that providing participants with additional trials to practice an atypical study condition like singing may have helped facilitate the singing superiority effect. However, an exploratory analysis of data from Whitridge (2022) suggests that mnemonic benefits for singing are no more pronounced for items occurring late in the study list (i.e., when participants have had more practice) relative to early items.

Additionally, Quinlan and Taylor's (2013) recognition test used correctable yes/no judgements, whereas Whitridge (2022) employed six-point confidence judgements. Some researchers have suggested that increasing the number of points available in a confidence judgement increases decision noise and can thereby decrease the precision of the estimate produced by the judgement (e.g., Benjamin et al., 2013). However, Whitridge (2022) reasoned that this would not likely obfuscate a singing superiority effect, given that decision noise should impact judgements equally in all conditions. Nonetheless, all discrepancies in the study and test phases were adjusted to match Quinlan and Taylor (2013) in the present experiment.

One further difference between Quinlan and Taylor (2013) and Whitridge (2022) that might be of particular importance is the list of words used at study. The latter study used a

slightly modified version of the stimuli from Ozubko et al. (2020), which was generated from a different corpus than that used by Quinlan and Taylor (2013). This led me to speculate that item characteristics might have modulated the singing superiority effect. For example, being instructed to produce a monosyllabic item (e.g., *dog*) via singing might lead participants to sing the item using only one distinct pitch. On the other hand, participants might be more inclined to sing multisyllabic items using a variety of different pitches because singing with variation is more intuitive for these items. Importantly, it is possible that additional variation within singing could render the associated production trace more distinctive (see Section 5.3 for further discussion). Thus, the production effect for singing could be larger for study lists that consist of more items that encourage varied singing (e.g., items with a greater number of syllables). Furthermore, evidence suggests that processing of melodic information differs for consonant and vowel sounds, with the latter being processed more independently from melody relative to the latter (e.g., Kolinksy et al., 2009). It is reasonable to speculate, then, that singing words with more vowel sounds might help to bind the record of distinctive melodic processing to the production trace; if this renders distinctive melodic information more accessible at test, the magnitude of the production effect for singing might increase. Given the potential importance of item characteristics, I selected stimuli that were as similar as possible to those used by Quinlan and Taylor (2013), although the exact list used in that study could not be obtained.

The present experiment updated the materials and procedure of Whitridge (2022, Experiment 2) to replicate Quinlan and Taylor (2013, Experiment 2) as exactly as possible. To this end, I modified the stimuli, pre-study phase, study phase and test phase from Whitridge (2022) in accordance with the methods reported by Quinlan and Taylor (2013); as in Whitridge (2022), foil matching at test was included as a between-subject manipulation. If evidence for a

singing superiority effect were to emerge under these conditions, it would suggest that the relative superiority of singing over reading aloud is likely facilitated by hidden moderators related to study design. In this experiment, I expected to observe evidence for a production effect on sensitivity (*d'*) for both reading aloud and singing relative to reading silently. However, given that Whitridge (2022) observed evidence against the superiority of singing and that methodological changes made to the present experiment are minor, I expected to observe no evidence for a difference in sensitivity between the two production modalities.

2.2 Method

2.2.1 Participants

Participants in Experiment 1 consisted of 102 undergraduate students (N = 51 matched) from The University of Southern Mississippi who took part in the experiment in exchange for partial course credit.

2.2.2 Stimuli and Apparatus

Stimuli were selected by using the Paivio et al. (2009) word generator (http://euclid.psych.yorku.ca/shiny/Paivio/) to create a list of 240 words. Parameters for word generation were selected such that the characteristics of the generated stimuli were as close as possible to the characteristics reported by Quinlan and Taylor (2013); the corpus from which the words were selected was also the same as that used by Quinlan and Taylor. All words were nouns between three and seven letters in length, with a mean 5.20 letters (SD = 1.13) and a mean 1.54 syllables (SD = 0.63). The stimuli had a mean concreteness rating of 3.93 (SD = 1.10) and a mean SUBTLEX frequency score (Brysbaert & New, 2009) of 103.18 (SD = 194.34).

Each participant saw all possible words over the course of the experiment. Half the words (120 items) appeared in the study phase and were randomized between the three study conditions

for each participant (i.e., 40 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, white or blue; color assignments were counterbalanced across participants. All words were displayed in their assigned color at study and at test. The remainder of the words (120 items) appeared only as "new" foils at test. For the matched group, the color assignment for foils was randomized across the three possible assignments, such that an equal number of foils appeared in each possible color. 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 experiment was coded in PsychoPy (version 3.4.2; Peirce et al., 2019) and presented via a 20-inch color monitor attached to a computer running Windows 10.

2.2.3 Procedure

The experiment consisted of a study phase and a test phase. Prior to the study phase, a researcher gave participants verbal instructions that were later reiterated within the experimental program. Participants were informed that they would see words presented one at a time in one of three colors (red, white, or blue) and that the color indicated how the words should be studied. The specific color assignment for each study condition was counterbalanced between participants. For the read silently condition, participants were instructed to read the words silently without any vocalization or mouth movement. For the read aloud condition, participants were instructed to read the words aloud in a normal voice. For the sing condition, participants were instructed to sing the words aloud as they would sing in any other context (e.g., in the car or in the shower). 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. Prior to beginning the study phase, participants were told that they would

first complete a familiarization phase followed by a practice phase to make sure they understood each study condition.

Familiarization Phase. In the familiarization phase, participants were presented with 15 trials. Participants saw 5 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, 5 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. After completing the practice phase, participants moved on to the study phase.

Study Phase. During the study phase, participants were presented with a series of 120 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 (40 items each). Each trial began with a 500 ms blank screen and then the word at center for 2000 ms. Participants were supervised by an experimenter throughout the study phase. After all practice 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, 120 of which were "old" words seen in the study phase and 120 of which were "new" foil words. For the matched group, "old" words were presented in the same color to which they were assigned

at study (e.g., a word presented in blue at study was presented in blue at test). Foils were randomly intermixed between the three color assignments such that 40 foils were presented in each possible color. For the unmatched group, all items were presented in yellow at test; 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 judgement as to whether the word was previously studied (i.e., "old"). Judgements were made using a textbox that appeared below the word, in which participants could respond by pressing either the "Y" key (yes) or the "N" key (no). Participants could correct their responses using the backspace key. When they were ready, participants submitted their response to each trial using the space bar. After a response was submitted, the next word was presented at center; this repeated until participants had completed all 240 trials.

2.2.4 Statistical Approach

Several approaches were used to analyze the data from Experiment 1. My primary analysis used multilevel probit regression to estimate signal detection parameters for each study condition. Subsequently, I fit exploratory diffusion models to model response times whilst also accounting for accuracy. Finally, I conducted exploratory analyses of serial position effects in the data using generalized additive mixed models (GAMMs). Below, I provide a detailed overview of each approach.

2.2.4.1 Signal Detection Analysis

Rather than analyzing raw or corrected hit rates, I opted to estimate sensitivity (*d'*) and response bias (*C*), parameters derived from *signal detection theory* as applied to recognition memory (Egan, 1958). According to this model, participants use a *decision criterion* to determine whether a test item is old (i.e., previously studied) or new (i.e., a foil item). At test, each item

elicits some degree of familiarity, which is compared to the participant's decision criterion: If the familiarity elicited by a given item exceeds the criterion, the participant will make an "old" response. If the familiarity elicited is below the criterion, the participant will instead make a "new" response. Both studied and unstudied items are expected to elicit some degree of familiarity, although the degree to which a given stimulus is familiar is assumed to vary across trials. Thus, distributions of familiarity arise for both studied items (i.e., a "signal" distribution) and unstudied items (i.e., a "noise" distribution). However, participants are expected to make more "old" responses for studied relative to unstudied items, given participants' recent exposure to the former. Thus, the means of the signal and noise distributions are assumed to differ, with the mean of the former being higher than the latter under typical circumstances. My parameters of interest can be derived from this general model: *C* reflects the threshold of familiarity at which participants will make an "old" response (i.e., the decision criterion), whereas *d'* reflects the distance between the signal and noise distributions. In other words, *C* quantifies the participant's propensity to make "old" responses regardless of whether an item was previously studied (with higher values reflecting more conservative response bias), and *d*' quantifies the participant's propensity to discriminate successfully between old and new items (with higher values reflecting better discriminability; for a detailed overview of signal detection theory in the context of recognition memory, see Macmillan & Creelman, 2005; see also, Banks, 1970; Stanislaw & Todorov, 1999).

My decision to adopt a signal detection approach throughout the present thesis was motivated by two factors. First, a large body of research on recognition memory suggests that analyses using only raw or corrected hit rates do not adequately quantify participants' true capacity for discrimination (e.g., Rouder et al., 2007; Stanislaw & Todorov, 1999; see Macmillan

& Creelman, 2005, for a detailed discussion). Secondly, investigations of the production effect have demonstrated a tendency towards more liberal response bias for produced (relative to unproduced) items. Accordingly, these findings argue for the superiority of signal detection analysis for interpretation of the production effect on the basis that ignoring response bias may overestimate the size of the effect and bias inference (e.g., Fawcett et al., 2012, 2023).

With this framework in mind, I opted to use Bayesian probit regression to estimate *d'* and *C* in a multilevel context. Although this methodology has seldom been used within the production literature (e.g., Fawcett & Ozubko, 2016; Zormpa et al., 2019), the approach possesses several advantages over more conventional analyses. First, the primary dependent measure in the present experiment was binary (i.e., old/new responses). Typically, studies of the production effect have collapsed binary trial data into aggregate hit and false alarm rates for each participant, which produces response variables that are appropriate for conventional statistical approaches such as analysis of variance (ANOVA; e.g., Gathercole & Conway, 1988; MacLeod et al., 2010). However, aggregating trial data into proportions often produces data that are (1) heteroscedastic, thereby violating a core assumption made by ANOVA, and (2) constrained between zero and one, which typical ANOVA models cannot account for; these problems have been shown to lead to increases in Type I error rates and such procedures are thereby liable to detect spurious effects (e.g., Dixon, 2008; Jaeger, 2008). While these issues can be addressed to some extent by using conventional procedures to calculate *d'* and *C* (e.g., Stanislaw & Todorov, 1999), this approach still necessitates that data be aggregated into proportions prior to applying transformations. This is problematic, as collapsing data across items and participants fails to capture variability that arises across these parameters and can systematically bias inference (Baayen et al., 2002; Gelman & Hill, 2006; Rouder et al., 2007). Given that 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, this variability cannot be safely ignored in the present experiment.

However, these problems can be mitigated by using generalized linear mixed models (GLMMs) to estimate signal detection parameters directly (Rouder & Lu, 2005; Rouder et al., 2007; Wright et al., 2009). Approaching categorical data analysis using logistic or probit regression models does not necessitate that data be collapsed across trials and avoids systematic inflation in error rates that can arise due to heteroscedasticity in ANOVA models (Dixon, 2008; Jaeger, 2008). Furthermore, GLMMs permit the inclusion of random effects, which account for variability that is expected to arise at the level of the item or the participant (Baayen et al., 2002; Barr et al., 2013; Gelman & Hill, 2006). My decision to implement these models using a fully Bayesian approach was motivated firstly by the capacity of Bayesian models to incorporate regularizing prior knowledge: Because both my dependent measures of interest possess lower and upper bounds, and it is reasonable to establish an *a priori* range in which these parameters can be expected to fall. Furthermore, 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). The probit models implemented herein deviate slightly from comparable approaches within the production literature, which have utilized

logistic regression to estimate analogous parameters (e.g., Fawcett & Ozubko, 2016; for a comparable Frequentist approach, see Zormpa et al., 2019; see also, Fawcett et al., 2016). However, logistic and probit approaches to multilevel 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, I opted for probit regression over logistic regression to ensure that my parameter estimates could be easily interpreted by readers familiar with signal detection theory.

The models reported herein were parameterized in accordance with the general structure for generalized linear signal detection models outlined by DeCarlo (1998; see also, Macmillan & Creelman, 2005; 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

$$H = \Phi(d'/2 - C)$$

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

$$F = \Phi(-d'/2 - C)$$

where Φ denotes the normal cumulative distribution function (Macmillan & Creelman, 2005). To estimate the probability, *p*, of an "old" response for a given trial event, *k*, let *old* denote a binary classification variable that is coded as either ½ or -½. The latter coding indicates that the target item for a given trial was previously studied, whereas the former coding indicates that the target is a foil. The probability of an "old" response on the *k*th trial is then given by

$$p_k = \Phi(d' old_k - C)$$

Thus, the probability of a hit is calculated when *old* = ½, and the probability of a false alarm is calculated when *old* = -½. All models were parameterized using the general structure given by this equation, albeit also including fixed and random effects as applicable for a given design. This parameterization permitted *d'* and *C* to be estimated directly, with fixed and random effects applied independently to each parameter. Using this approach, I removed the model intercept and computed slopes that produced estimates of *d'* and *C* for all possible combinations of the fixed effects.

Each model included fixed effects for condition (sing, aloud, silent) and group (matched, unmatched) applied to both *d'* and *C*. I assumed that the impact of the fixed effects would vary across participants and items, given that 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, I included random slopes to permit item-level variation in the impact of the fixed effects of both group and condition. For participant-level effects, however, the 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 I removed the intercept from the models in all cases, I also modeled correlations between random slopes reflecting my assumption of baseline variability in *d'* and *C* across participants and items.[11] While not all model terms corresponding to the random effects are reported in-text, an overview of these estimates is provided for each signal detection model reported below.

[11] 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.

For each model, I applied uninformative, mildly regularizing priors. These priors were specified to reflect my 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) and the lower and upper boundaries that *d'* and *C* can reasonably assume. For random effects, my priors were calibrated to beliefs that the standard deviation for any given clustering variable across these parameters should fall between 0 and 2. Finally, where applicable, I also applied regularizing *lkj* priors to correlations between random effects with a scale of 4 (Lewandowski et al., 2009). These priors essentially regularize correlation coefficients in a manner similar to the priors for parameters reported above, placing greater weight upon coefficients close to zero—but nonetheless allowing for reasonably high estimates (for further discussion, see McElreath, 2018).

All models were fit using 8 independent sampling chains of 15000 iterations each. However, because each chain requires a warm-up period to converge to the posterior, the first 7500 samples were discarded, as is standard practice (see Kruschke, 2014, and McElreath, 2018, for technical discussion of sampling procedures used in Bayesian parameter estimation). Thus, the models included 60000 post-warmup draws in total. Model convergence was assessed using visual inspection of the chains and R-hat statistics, which were less than 1.01 in all cases, indicating that the models converged (Gelman & Hill, 2006; Kruschke, 2010). Further, inspection of the chains showed that both bulk and tail effective sample sizes were equal to or greater than 20000 for all estimates reported in-text. Although sampling chains for some item-level correlation terms were less efficient, effective sample sizes for these coefficients were nonetheless greater than 7000 in all cases.

For each model, I 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 coefficient and reflect raw differences in each parameter. Alongside these parameters, I 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 my analyses of response bias, the colour matching procedure allowed me 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 (i.e., by randomly distributing foil items across conditions). As a result, condition-specific estimates of this parameter for the unmatched group capture differences in hit rates but not false alarm rates. Thus, estimates of response bias can be meaningfully interpreted only for the matched group; my discussion of this parameter thereby focuses solely on the matched group. With this limitation in mind, I also report median posterior estimates for *C* by condition and contrasts between conditions; 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.

2.2.4.2 Diffusion Models

An alternative approach to analyzing data derived from binary decision-making tasks (e.g., recognition memory) is the *diffusion model* (Ratcliff, 1978; for reviews, see Voss et al., 2013; Wagenmakers, 2009). Broadly, this model assumes that after a to-be-judged stimulus is presented, a process of evidence accumulation is initiated. This accumulation process is noisy, but the evidence will eventually cross one of two possible thresholds corresponding to the two available decisions, prompting a response. In a simple recognition paradigm, for example, participants encounter words at test and are asked to decide whether the words were previously studied or not. Thus, in this example, the two boundaries represent decisions corresponding to "old" and "new" responses, respectively. According to the diffusion model, when a participant encounters a test word, the participant begins to accumulate evidence with respect to whether the word matches one previously studied. As evidence accumulates, the participant will "drift" towards one of the boundaries; once one of the boundaries is reached, the participant will decide whether they recognize the word (Ratcliff et al., 2004).

From this model, several key parameters of interest can be derived. The central parameter of the diffusion model is *drift rate*, which refers to the rate at which evidence toward either boundary is accumulated. As applied to recognition memory, this parameter reflects the quality of the agreement between the test item and the participant's memory; in other words, higher drift rate equates to accumulation of evidence that is faster, more accurate, or both. The second parameter of interest is *boundary separation*, which refers to the distance between the two decision boundaries. This parameter quantifies the amount of evidence required to make a

decision, with a larger boundary separation indicating that more information is necessitated. Diffusion models also include a parameter corresponding to *starting point* (sometimes referred to as *bias*, e.g., Wiecki et al., 2013), which reflects the point at which the process of evidence accumulation begins; this parameter quantifies bias towards either boundary. Finally, *non-decision time* quantifies the amount of time participants spend encoding the stimulus prior to initiation of the decision process (Ratcliff & Rouder, 1998; Ratcliff et al., 2004). Although more complex variants of the diffusion model including additional parameters exist (e.g., parameters corresponding to trial-by-trial variability in drift rate, starting point, and non-decision time; Ratcliff & Rouder, 1998), the four-parameter version detailed above is commonly employed for recognition memory and is easily implemented within a Bayesian framework (Bürkner, 2021; Wiecki et al., 2013); as such, I chose to conduct my exploratory analyses using the four-parameter variant.

My decision to fit exploratory diffusion models to this data was motivated firstly by the absence of previous efforts to model production data in this manner: Although the diffusion model has been applied in a wide variety of cognitive tasks that involve binary decision-making (see, e.g., Voss et al., 2013; Wagenmakers, 2009), no published study within the production literature has adopted this approach. Further, diffusion models offer several advantages over conventional approaches to analysis of recognition data. For example, response times and the response variable of interest (e.g., and old/new response) are typically modeled separately. However, this approach is limited insofar as it generally does not account for the full distribution of response times (focusing instead on the mean), nor speed-accuracy trade-offs. On the other hand, the diffusion approach permits response times to be modeled whilst also accounting for a response variable; the approach is thus able to capture variation in each as well as any number of

interactions between the two variables. Furthermore, the parameters derived from the model allow for the quantification of latent cognitive processes that are ignored in conventional analyses; thus, diffusion modeling can provide a more complete picture of the decision process (Ratcliff & Rouder, 1998; Ratcliff et al., 2004; Wagenmakers, 2009).

With respect to my implementations of this model, an overview of the general parameterization and mathematical assumptions underlying diffusion models is far beyond the scope of the present thesis. Broadly, the diffusion models I implemented modeled response times and response accuracy as a joint process and produced estimates corresponding to drift rate, boundary separation, starting point, and non-decision time (for detailed overviews, see Feltgen & Daunizeau, 2021; Ratcliff, 1978; Wagenmakers, 2009; for a tutorial, see Singmann, 2017). I specified the models using a non-linear formula that allowed fixed and random effects to be applied separately to each parameter. Importantly, approaches to diffusion modeling typically assume that stimulus characteristics which are not known to participants prior to the start of each trial can only affect drift rate; this is because all other parameters are thought to be fixed prior to encountering the stimulus (see, e.g., Feltgen & Daunizeau, 2021). Accordingly, fixed effects corresponding to stimulus characteristics that could not be known before the stimulus was presented were applied only to drift rate. Thus, for the present experiment, the drift rate parameter included fixed effects for condition, "old" status (i.e., a categorical classification variable indicating whether an item was previously studied) and group, whereas the other parameters included only a fixed effect for group. The drift rate parameter also included a random slope that allowed for variability in the effects of condition, "old" status and group across items, as well as slopes that permitted variability in the former two fixed effects across participants. All other parameters included random slopes corresponding to item-level variability

across groups and random intercepts permitting baseline variability across participants. Due to issues with convergence, correlations between random effects were not included in the final models.

Initially, I modeled the data using two approaches that differed slightly: I first modeled the yes/no responses irrespective of correctness and then modeled accuracy (i.e., correct identification of an item as old or new). For the former approach, the models were implemented as described above. For the latter approach, however, starting point was fixed at the midpoint (i.e., no bias towards either decision), resulting in a three-parameter diffusion model that is standard for models of correctness (Voss et al., 2015; see also, Bürkner, 2021). Thus, for the latter approach, no estimates corresponding to starting point were produced. However, the four-parameter models did not converge as well as the simpler models and moreover, estimates of starting point produced by these models were not meaningfully different from the fixed bias specified for the models of accuracy (i.e., the credible intervals corresponding to estimates of this parameter contained 0.5); the inferences derived from each approach were also identical. Accordingly, I opt to report only the results of the simpler models herein.

In either case, I first prepared the data by removing all trials for which participants self-corrected their decision. To elaborate, the recognition test employed in this paradigm was atypical insofar as participants were able to correct their response before pressing another key to submit the response. Typically, diffusion models are employed for fast decisions (e.g., 1000 – 1500 ms on average; Ratcliff et al., 2004) and fitting such a model to the data collected in this experiment thereby presents a unique challenge. Although the opportunity to self-correct was available, the data showed that participants opted instead to immediately submit their initial decision on most trials. Drawing upon this finding, I opted to analyze response times for the

initial decision and discard trials for which multiple decisions were made; this resulted in the exclusion of < 5% of trials. Because no precedent exists for applying diffusion models to response times derived from such a paradigm, this approach is novel and is likely not without limitations. Accordingly, conclusions drawn from the exploratory diffusion models reported herein should be interpreted with caution. Subsequently, I discarded all trials for which response times were extreme by excluding observations that were below the 2.5% quantile and above the 97.5% quantile (see, e.g., Berger & Kiefer, 2021; see also, Ratcliff et al., 2018); this step was necessary to ensure that the models could successfully initialize and converge (Bürkner, 2021; Singmann, 2017).

For the diffusion models, I applied uninformative, mildly regularizing priors. Given that no published study has fit diffusion models to production data, my priors were broadly calibrated with respect to other studies of recognition memory (e.g., Ratcliff, 1978; Ratcliff et al., 2004) and the upper and lower boundaries that each parameter could hypothetically assume. For drift rate, I specified priors such that this parameter was reasonably expected to fall between -4 and 4. For boundary separation, nondecision time and starting point (where applicable), priors were specified such that each parameter was reasonably expected to fall between 0 and 2. For random effects, my priors were calibrated to beliefs that the standard deviation for any given clustering variable across these parameters should fall between 0 and 2. Finally, I also applied regularizing *lkj* priors to correlations between random effects with a scale of 3 (Lewandowski et al., 2009; for further discussion, see McElreath, 2018).

The diffusion models were fit using 6 independent sampling chains of 5000 iterations each, with a warmup period of 2500 iterations. Thus, the models included 15000 post-warmup draws in total. Model convergence was assessed using visual inspection of the chains and R-hat

statistics, which were less than 1.01 in all cases, indicating that the models converged (Gelman & Hill, 2006; Kruschke, 2010). Further, both bulk and tail effective sample sizes were equal to or greater than 1000 for all estimates reported in-text. For each model, I report the median posterior estimate corresponding to each parameter of interest alongside the 95% HDI.

2.2.4.3 Analyses of Serial Position

As outlined previously in Section 1.1 of the present thesis, recent research efforts by Saint-Aubin and colleagues (2021; Cyr et al., 2022; Gionet et al., 2022, in press) have provided evidence for an interaction between production and serial position in certain paradigms: In pure-list tests of recall, a reverse production effect (i.e., silent > aloud) arises for early positions, whereas a typical production advantage emerges for late positions. To expand upon my earlier discussion, this interaction has been explained with reference to a *revised feature model* (RFM), adapted from Nairne's (1990) feature model. According to this model, production interferes with rehearsal and can thereby hinder the encoding and maintenance of item features. In pure lists, items read silently are not subject to production-related interference (i.e., interference arising due to production-related sensory processes) because no items are produced; thus, these items are well-rehearsed at both early and late serial positions. On other hand, produced items in pure lists are subject to varying degrees of interference depending upon their serial position. For early positions, substantial interference arises and hinders rehearsal, whereas late items are subject to relatively less interference because few items are produced subsequent to these positions. Accordingly, a reverse production effect arises for early items because silent items are better rehearsed relative to aloud items, while a typical production advantage occurs for late positions (for discussion and computational implementation of the RFM, see Saint-Aubin et al., 2021). The notion that such an

interaction occurs in pure-list recall paradigms has received considerable support from recent reviews (Gionet et al., 2022) and meta-analyses (Fawcett et al., 2023).

However, evidence for an interaction between production and serial position in tests of recognition is mixed. One meta-analytic model of serial position effects in recognition found the production effect to be credible for both early and late positions, albeit with a numerical trend favoring a larger advantage for the latter (Fawcett et al., 2023). In contrast to these findings, however, a recent series of empirical investigations by Gionet et al. (in press) found no evidence for an interaction between production and serial position in recognition. These authors speculated that the interaction predicted by the RFM might have failed to emerge due to the greater list length typical of recognition paradigms: Participants rehearse a lower proportion of items as list length increases, reducing the cost of production-related interference for early items. Given this discrepancy in findings, I speculated that the statistical approach employed by Gionet et al. (in press) could have obfuscated the critical interaction. Whereas those authors fit single- and multilevel linear models to the data, earlier investigations of serial position and production (e.g., Cyr et al., 2022; Gionet et al., 2022) appear to show nonlinear relationships between the variables that would not likely be captured by typical approaches (see, e.g., Wood, 2017, for a detailed discussion). On this basis, I opted to investigate serial position effects in my own data using an approach designed to capture the possibility of a nonlinear relationship.

As a secondary objective, I implemented models of serial position to potentially capture practice effects within my data. Earlier work by Wakeham-Lewis et al. (2022) observed non-significant trends hinting at an increase in the size of the production effect for character voices – but not for reading aloud – for later list positions. Given that both singing and reading in character voices constitute unusual production modalities with which participants may be unfamiliar (see

Sections 1.3 and 5.2 of the present thesis for further discussion), it is possible that similar patterns could emerge in my data. If this were the case, it would suggest that the singing superiority effect is driven in part by experience with the modality (but see Quinlan & Taylor, 2019).

With respect to the precise implementation of my serial position analyses, a detailed overview of the mathematical theory underlying GAMMs lies outside the scope of this thesis; interested readers are directed to Wood (2017). In simple terms, additive models allow for functions describing patterns in the data to be constructed as the weighted sum of any number of simpler basis functions. Thus, a complex, nonlinear pattern that could not be captured by a single polynomial function can instead be captured by a piecewise function composed of multiple components. Practically speaking, this approach allowed me to fit probit models inclusive of serial position effects that were capable of capturing complex nonlinear trends in my data—although models fit using this approach are penalized by smoothing parameters for departures from linearity (i.e., curves that are "wigglier" are penalized more than smoother curves) to mitigate the risk of overfitting.

For reasons described above, I opted to fit signal detection models (rather than models of raw or corrected hit rates) inclusive of fixed and random effects corresponding to serial position (i.e., the order in which items appeared during the study phase). Because foil items do not have a meaningful position within the study phase order, I first randomly (and arbitrarily) split foil items across serial positions, thus assuming a common false alarm rate for each position. I then modeled standardized serial position first as a nonlinear effect and subsequently as a linear effect; this approach was taken to allow for the possibility that serial position effects could be linear. For either approach, the parameterization of the models was similar to the probit models of old/new recognition detailed above, albeit inclusive of fixed and random effects corresponding to serial

position applied to *d*' (but not *C*).[12] For the linear models, I included fixed effects for serial position across conditions as well as random slopes for serial position across participants; the structure of the nonlinear models was identical, albeit instead including fixed and random smooth terms that permitted the estimation of nonlinear serial position effects. Initially, I intended to include random effects such that participant-level random curves for serial position were permitted to vary across conditions. However, nonlinear implementations of the models inclusive of these effects converged poorly; accordingly, my primary linear and nonlinear models proceeded instead with a simplified random effects structure excluding only that term. I opted not to model interactions between serial position and group on the basis that differences related to color matching would be expected to emerge primarily in false alarm rates, which could not be meaningfully captured by the modeling approach taken here. Further, I chose to exclude item-level random slopes and curves for serial position for two reasons: First, because there is no obvious theoretical basis from which to suggest that item-level variation in serial position effects should emerge; and second, because fixed and random effects corresponding to serial position should already account intrinsically for some degree of intertrial variability.

For the models of serial position, priors on *d'* and *C* were largley identical to those described for the probit models above. For these models, however, I also placed uninformative, mildly regularizing priors on the effect of serial position on *d'*. For the linear models, these priors were specified such that serial position-related changes in *d'* in any given condition were expected to fall between -1 and 1; identical priors were placed on the linear trend of serial position in the nonlinear models. For the nonlinear models, I also placed priors on the variance of

[12] Given that my random assignment of foil items to study positions meant that estimates of *C* could not be meaningfully interpreted, I opted to apply fixed and random effects corresponding to serial position exclusively to *d'*. In support of this decision, models inclusive of serial position terms applied to both parameters produced identical inferences to the models in text, but convergence was superior for the latter.

the smooth parameters (i.e., the "wiggliness" of the curves) such that these coefficients were expected to fall between 0 and 4. In all cases, priors were calibrated to be largely uninformative and allow for a wide – but reasonable – range of possible effects.

Both the linear and nonlinear models were fit using the *brms* package (Bürkner, 2017) in *R* (R Core Team, 2020); the latter approach additionally made use of functions imported from the *mgcv* package (Wood, 2011). With further respect to the nonlinear analyses, these models were fit using the default number of 10 knots (i.e., the number of points between the component basis functions). This approach was taken because I had no strong theoretical basis from which to describe the proximate curve expected to arise from the data. Although it is very likely that 10 knots was greater than the number of knots necessitated by my data, no issues related to overfitting should arise because wigglier curves are penalized accordingly. Due to the extensive computational cost of fitting additive mixed models using a Bayesian approach, all models were fit using 6 independent chains of 5000 iterations each, with a warm-up period of 2500 iterations. Thus, each model included a total of 15000 post-warmup draws. Model convergence was assessed using R-hat statistics, which were less than 1.01 in all cases, indicating that all models converged (Gelman & Hill, 2006; Kruschke, 2010). Effective sample size was greater than 6000 for all estimates reported in this supplement and greater than 10000 in most cases. Chains corresponding to some random effects were less efficient, but effective sample size for these coefficients was greater than 2000 in all cases and greater than 5000 in most cases.

2.3 Results and Discussion

Table 2.1 shows means and standard deviations for all primary dependent measures as a function of condition and group. Table 2.2 shows means and standard deviations for response times as a function of condition, group, and item type.

Table 2.1

Mean Proportion and Standard Deviation of the Mean for Hit Rates, Corresponding False Alarm Rates, Sensitivity (d'), and Response Bias (C) as a Function of Condition and Group

Condition	Hits	FAs	*d'*	*C*
Matched				
Sing	.68 (.16)	.17 (.15)	1.65 (.63)	.30 (.50)
Aloud	.68 (.16)	.21 (.16)	1.46 (.56)	.21 (.50)
Silent	.48 (.20)	.24 (.17)	.78 (.41)	.46 (.58)
Unmatched				
Sing	.69 (.15)	.22 (.15)	1.44 (.55)	.17 (.44)
Aloud	.68 (.15)	.23 (.14)	1.33 (.55)	.15 (.39)
Silent	.45 (.18)	.23 (.15)	.70 (.45)	.51 (.48)

Note. Sensitivity (*d'*) and response bias (*C*) were calculated conventionally (rather than estimated via Bayesian probit regression).

Table 2.2

Mean and Standard Deviation of the Mean for Response Times (RTs) in Seconds as a Function of Condition, Group, and Item Type

Condition	Item Type	
	Old	New
Matched		
Sing	1.49 (.83)	1.47 (.80)
Aloud	1.43 (.76)	1.51 (.83)
Silent	1.51 (.82)	1.48 (.82)
Unmatched		
Sing	1.31 (.74)	1.37 (.77)
Aloud	1.33 (.73)	1.32 (.72)
Silent	1.41 (.79)	1.36 (.80)

Note. Descriptive statistics for RTs were calculated after the removal of outlier trials, as outlined above.

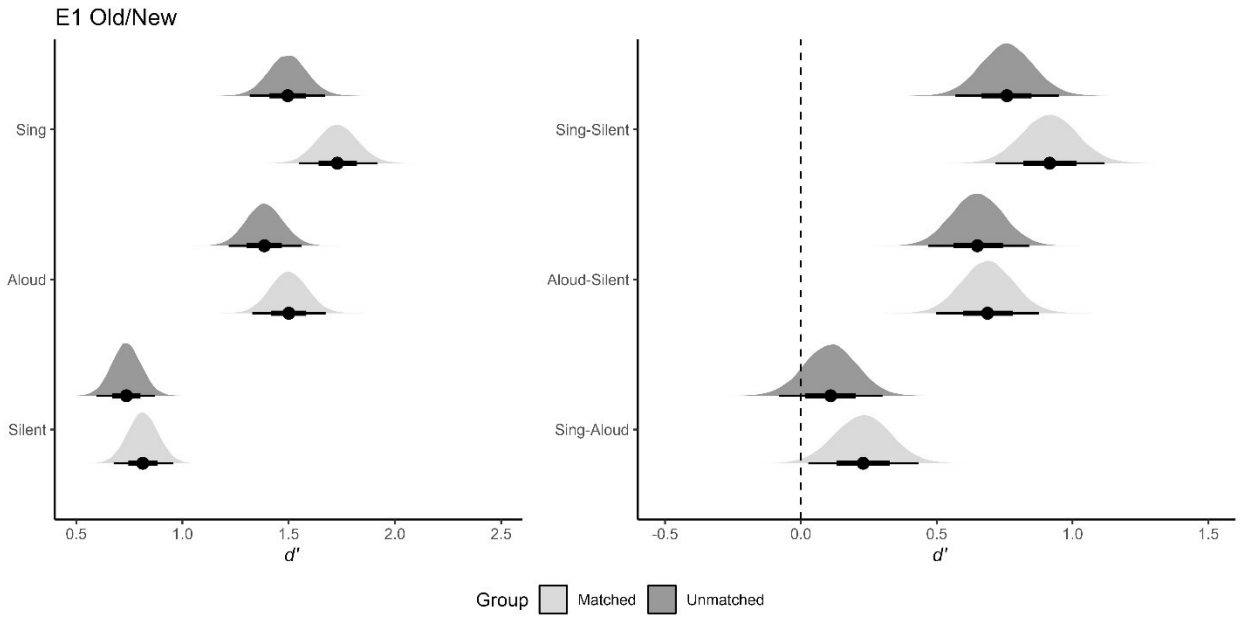
2.3.1 Signal Detection Analysis

I applied a multilevel probit regression to old/new responses, which was parameterized as described above. Below, I report the results for each parameter of interest estimated by the model (i.e., *d'* and *C*).

Sensitivity. Figure 2.1 shows posterior estimates for sensitivity by condition and group and contrasts between conditions. As depicted in Figure 2.1, the production effects for either modality were credible across groups. These findings are unsurprising and align well with prior studies that have observed robust production effects for singing (e.g., Quinlan & Taylor, 2013, 2019; Whitridge, 2022). Importantly, however, a small but credible singing superiority effect emerged in the matched group (estimate = 0.11, $HDI_{95\%}$ = 0.01 – 0.21), although this pattern did not occur for the unmatched group (estimate = 0.02, $HDI_{95\%}$ = -0.07 – 0.12). Contrary to my hypotheses, then, it appears that singing *does* result in a larger production effect relative to reading aloud, at least under certain circumstances. However, Whitridge (2022; see also, Hassall et al., 2016) previously failed to detect the effect across multiple experiments, with the effect emerging only in the present experiment when the design of the study was a near-exact replication of Quinlan and Taylor (2013; Experiment 2). Accordingly, my observation of a credible singing superiority effect comes with the important caveat that the effect appears to be driven – at least in part – by aspects of study design. In tentative support of this notion, a numerical trend favored a larger singing superiority effect in the matched group relative to the unmatched group (estimate = 0.08, $HDI_{95\%}$ = -0.05 – 0.21), a pattern which was also observed in Whitridge (2022; Experiment 3). Taken together, these findings hint strongly at the possibility of an interaction between singing and color matching. I further explored this point in Chapter 4 of the present thesis.

Figure 2.1

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) 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.

As outlined previously, the design of Experiment 1 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 > 0.31), indicating variability in the impact that study modality had on sensitivity across individuals. Participant-level correlations between the aloud/sing and sing/silent conditions were moderate and positive (estimates > 0.48), suggesting overall that participants with higher sensitivity in one condition also exhibited higher sensitivity in other conditions.

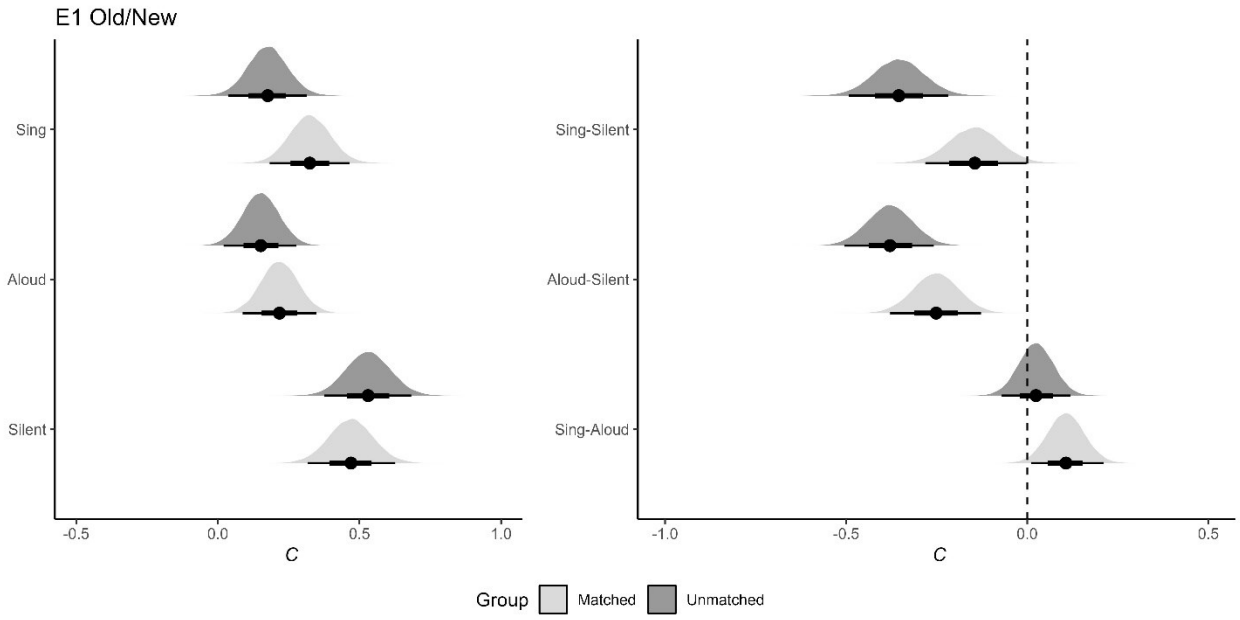
Item-level random slopes corresponding to the effects of condition and group were also informative in all cases (estimates > 0.41 for the matched group and > 0.35 for the unmatched group). While correlation terms corresponding to item-level effects were less informative than those corresponding to participant-level effects, the majority of the estimates were credible and positive, suggesting some degree of baseline variation in sensitivity across items.

Response Bias. Figure 2.2 shows posterior estimates for response bias by condition and group and contrasts between conditions. As depicted in Figure 2.2, response bias across groups 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 = 0.11, $HDI_{95\%}$ = 0.01 – 0.21). This trend is novel, with earlier studies observing similar bias for sing and aloud items (Quinlan & 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 previously), 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 > 0.25). Participant-level correlations between all conditions were moderate to strong and positive (estimates > 0.63).

Figure 2.2

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) 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.

Item-level random slopes corresponding to the effects of condition and group were also informative in all cases and similar across groups (estimates > 0.25 for the matched group and > 0.24 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 > 0.41 for the unmatched group), indicating baseline variation in response bias across items. Overall, these trends are generally consistent with those observed for the random effects on sensitivity.

2.3.2 Diffusion Models

In addition to a standard signal detection model, I applied a multilevel diffusion model to accuracy, which was parameterized as described previously. I computed a single value of drift

rate for each condition by adding the posterior of the parameter for old items to the posterior of the parameter for new items (see, e.g., Ratcliff et al., 2021, 2022). This approach allowed me to simultaneously account for drift rates for both old and new items, akin to a signal detection analysis.

For the matched group, a credible production effect on drift rate emerged for both the sing (difference = 0.78, $HDI_{95\%}$ = 0.61 – 0.96) and aloud conditions (difference = 0.65, $HDI_{95\%}$ = 0.46 – 0.83), indicating faster accumulation of evidence towards correct responses for produced relative to unproduced items. Additionally, a non-credible trend supported higher drift rate in the sing condition relative to the aloud condition (difference = 0.13, $HDI_{95\%}$ = -0.07 – 0.33). For the unmatched group, credible production effects again emerged for both sing (difference = 0.70, $HDI_{95\%}$ = 0.50 – 0.90) and aloud (difference = 0.67, $HDI_{95\%}$ = 0.46 – 0.88), but there was no trend toward a singing superiority effect (difference = 0.03, $HDI_{95\%}$ = -0.15 – 0.23). Generally, these results echo the findings I observed across my signal detection models, demonstrating robust production effects for either modality. Nonetheless, these findings are the first to extend the production effect to drift rate, providing evidence that the advantage persists when jointly accounting for both accuracy and response times.

As discussed above, constraints associated with diffusion models meant that boundary separation and nondecision time could not be separated by condition or "old" status and were thus separated only by group. With respect to boundary separation, a credible effect of group emerged such that boundary separation was higher for the matched group relative to the unmatched group (difference = 0.16, $HDI_{95\%}$ = 0.01 – 0.32), suggesting that participants require more evidence before making a decision when color matching is present. This notion could be interpreted as consistent with differential use of heuristic strategies across groups: Knowing the

way an item has been studied could lead participants to search for specific sensorimotor information not normally sought in typical paradigms. Changes in boundary separation are often indicative of a speed-accuracy trade-off, wherein higher boundary separation equates to better accuracy at the cost of slower responses (Ratcliff & McKoon, 2008). Thus, participants might leverage specific information and think more carefully prior to adjudication when oriented to cues via stimulus dimensions (see Sections 5.2 and 5.3 of the present thesis for further discussion of this hypothesis). Finally, there were no credible differences nor trends in nondecision time across groups (difference = 0.01, $HDI_{95\%}$ = -0.06 – 0.08)

2.3.3 Analyses of Serial Position

Below, I report the results of both the linear and nonlinear models of serial position as well as empirical comparisons between the linear and nonlinear models. Model comparison was undertaken using approximate leave-one-out cross-validation to compare expected log pointwise predictive densities (ELPDs) via the *loo* package (Vehtari et al., 2024; for further discussion, see Vehtari et al., 2017);[13] in all cases, model comparison diagnostics were good (all k estimates < 0.7). Because the nonlinear models were often favored over the linear models throughout the present thesis, I focus my discussion on the former. The RFM (Saint-Aubin et al., 2021) makes no specific predictions as to the effect serial position might have in mixed-list recognition paradigms; accordingly, my analysis of Experiment 1 has no immediate theoretical relevance to this framework and was undertaken primarily for completeness and to explore practice effects.

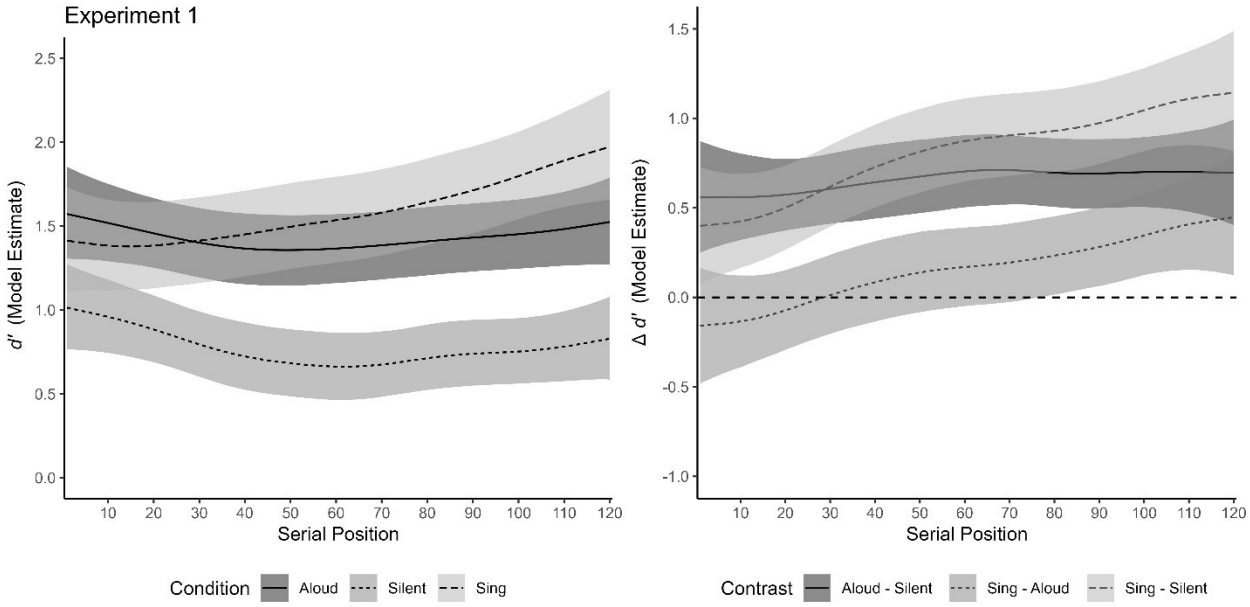
Given that additive models describe the data using a piecewise function made up of smaller components, it is difficult to derive straightforward inferences from smooth predictors. That is to

[13] For each comparison, I report the difference in ELPD between the candidate models (ΔELPD) and the standard error of the difference (ΔSE). These values can be interpreted such that models with lower ELPD generate better predictions. As a rule of thumb, I considered a model to have been supported if ΔELPD was greater than two times ΔSE (akin to a 95% confidence interval). However, readers should note that this convention is arbitrary.

say that unlike the linear models, the nonlinear models did not compute slopes for serial position that could be evaluated for credibility. To derive inferences from these analyses, then, I computed model predictions for *d'* across serial positions (excluding random effects) and compared model estimates for sensitivity at different positions. Figure 2.3 depicts conditional nonlinear model estimates for *d*' across serial positions and contrasts between conditions across serial positions (i.e., difference smooths for sing/aloud - silent and sing - aloud).

Figure 2.3

Nonlinear Model Estimates for d' as a Function of Condition and Serial Position



Note. The shaded region surrounding each curve depicts the 95% quantile interval of the estimate.

For the linear model, the slope for serial position narrowly reached credibility for the sing condition (estimate = 0.08, $HDI_{95\%}$ = 0.00 – 0.16), but not for the aloud or silent conditions. In this case, model comparison failed to provide definitive evidence, but marginally favored the

nonlinear model over the simpler linear model (ΔELPD = -4.5, ΔSE = 3.6) and definitively favored the nonlinear model over the more complex linear model (ΔELPD = -164.7, ΔSE = 17.9). To evaluate whether the inclusion of terms corresponding to serial position was beneficial, I also compared models inclusive of serial position effects to comparable models excluding these terms (i.e., parametrized identically to the signal detection analysis described above). For this comparison, the inclusive nonlinear model was marginally favored over the exclusive model (ΔELPD = -16.5, ΔSE = 11.4).

As shown in Figure 2.4, for the nonlinear model, the production effect was credible for either modality at all serial positions. Additionally, I computed planned contrasts comparing the initial 3 items to the final 3 items. These positions were selected to mimic those used in analyses by Fawcett et al. (2023; see also, Gionet et al., in press), who selected these positions arbitrarily. In this case, the planned contrasts provided credible evidence for a larger production effect for late items relative to early items in the sing condition (difference = 0.74, $HDI_{95\%}$ = 0.43 – 1.06), but only a non-credible trend emerged for reading aloud (difference = 0.14, $HDI_{95\%}$ = -0.18 – 0.45). Interestingly, this trend extended to the singing superiority effect (difference = 0.60, $HDI_{95\%}$ = 0.29 – 0.91), which was credible from position 77 onwards.

My serial position analyses of Experiment 1 hint strongly at a practice effect confined to the sing condition. Interestingly, this interaction emerged in an experiment which utilized both a practice phase and a longer study phase list (120 items versus 90 items), methodological features which were present in earlier studies that observed a singing superiority effect (e.g., Quinlan & Taylor, 2013, 2019) but not in earlier conceptual replications that failed to observe the effect (Whitridge, 2022). Further, the singing superiority effect peaked at the latest positions in the list, with positions this extreme not existing in earlier conceptual replications. While it could be the

case that the singing superiority effect emerges only when participants have had significant practice with the modality, I caution against definitively accepting this interpretation for several reasons. First, exploratory serial position analyses of data from Whitridge (2022) detected no such trend towards a larger singing superiority effect for later positions; in fact, the effect marginally decreased for later positions. Additionally, although the singing superiority effect in Experiment 1 peaked at the latest serial positions, the effect was credible from position 77 onwards. Given that conceptual replications reported in Whitridge (2022) used study lists consisting of more than 77 items (i.e., 90 items), a practice-related interaction would be expected to emerge prior to positions that were not present for the shorter lists. Furthermore, the amount of practice afforded by the additional pre-study phase in Experiment 1 was negligible: This phase consisted of 15 trials split equally between conditions, meaning that participants completed only five practice singing trials. Finally, earlier research has shown that experienced singers – who would be expected to be more practiced with the modality – do not show a singing superiority effect that is meaningfully different from those previously observed in typical samples (Quinlan & Taylor, 2019; see Section 5.3 of the present thesis for further discussion of practice- and experience-related effects).

Chapter 3: Experiment 2

3.1 Overview

In Experiment 2, I evaluated Quinlan and Taylor's (2019) claim that the production effect for singing occurs only in mixed-list, within-subject experimental designs. Quinlan and Taylor (2019) manipulated production (sing, aloud, silent) between-subjects to test a strength account of the singing superiority effect. Drawing on early research into the production effect (e.g., MacLeod et al., 2010; Ozubko & MacLeod, 2010), Quinlan and Taylor (2019) speculated that the singing superiority effect should persist between-subject if singing results in increased encoding strength

for memory traces relative to reading aloud. After failing to observe a production effect for singing altogether, the authors concluded that the benefit must be driven predominantly by distinctiveness.

However, Quinlan and Taylor (2019) also failed to observe a between-subject production effect for reading aloud. Recent meta-analytic evidence from Fawcett et al. (2023) suggests that the production effect is reliable in between-subject designs, but also that the benefit is generally much smaller in magnitude than its within-subject counterpart. Accordingly, failures to observe a between-subject effect (e.g., Hopkins & Edwards, 1972; MacLeod et al., 2010; Ozubko & MacLeod, 2010) might often be attributable to insufficient statistical power to reliably detect the effect. In the case of Quinlan and Taylor (2019, Experiment 4), the authors recruited 20 participants per group. Based on power analyses conducted by Fawcett et al. (2023), then, this experiment would have achieved less than 30% power to detect a typical between-subject production effect on sensitivity (i.e., $d = \sim 0.30$). Given that both the present investigation and Whitridge (2022) have largely suggested that the production effect for singing is very similar to that for reading aloud, it is reasonable to speculate that Quinlan and Taylor (2019) failed to observe a between-subject effect due to a lack of statistical power. Accordingly, there is no obvious reason that a well-powered experiment should not be able to detect a between-subject production effect for singing.

To test this hypothesis, the present investigation conceptually replicated Quinlan and Taylor (2019, Experiment 4). Fawcett et al. (2022) recommended a minimum sample size of 64 participants per group in order to ensure 80% power for detecting typical between-subject production effects on sensitivity. However, this experiment was conceptualized and implemented prior to the publication of Fawcett et al. (2023) and did not meet their recommendations for recruitment. Nonetheless, the sample size in the present experiment was more than double that of Quinlan and Taylor (2019, Experiment 4) and thereby had a much greater chance of detecting an

effect. In this experiment, I expected to observe evidence for a production effect on sensitivity for both singing and reading aloud relative to reading silently. If a between-subject production effect for singing is observed, this would suggest that the benefit for singing – much like that for reading aloud – persists across experimental designs and thereby cannot be explained solely by relative distinctiveness. Additionally, I expected to observe no evidence for a difference in sensitivity between singing and reading aloud.

As a secondary objective, the present experiment was the first investigation to examine the influences of recollection and familiarity processes on the between-subject production effect for singing; this was accomplished via the inclusion of recollect/familiar/neither judgements at test. Given that Fawcett and Ozubko (2016) found that the between-subject production effect was driven by familiarity (rather than recollection), I expected to observe evidence for a production effect on familiarity such that more "familiar" judgements are made for items that were either sung or read aloud at study relative to unproduced items. Finally, I did not expect to observe evidence for differences in "recollect" judgements between conditions.

3.2 Method

3.2.1 Participants

Participants in Experiment 2 consisted of 140 undergraduates 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 ($N = 43$).

3.2.2 Stimuli and Apparatus

For Experiment 2, stimuli were randomly selected from a pool of 360 words retrieved from the MRC Psycholinguistic Database (Coltheart, 1981) and previously used in both Ozubko et al. (2020) and Whitridge (2022). All words were nouns between 5 and 10 letters long, with a mean of 6.56 letters (*SD* = 1.54) and a mean 1.97 syllables (*SD* = 0.84). The stimuli had a mean concreteness rating of 3.42 (*SD* = 0.71) and a mean SUBTLEX frequency score (Brysbaert & New, 2009) of 82.01 (*SD* = 250.48).

Each participant was assigned a random subset of 180 words derived from the larger pool. Half the words (90 items) were studied by participants and were randomized between three possible color assignments (i.e., red, blue, and yellow) at study; this was implemented to maintain consistency with typical, mixed-list designs, wherein production is cued using color assignment. At test, these items were presented in white font. The remainder of the words (90 items) appeared only as "new" foils at test and were presented in white font. The experiment was coded in PsychoPy (version 2.3.2; Peirce et al., 2019) and presented via a 20-inch color monitor connected to a computer running Windows 10. All stimuli were presented in 14-point Arial font against a black background.

3.2.3 Procedure

Prior to the experiment, each participant was randomly assigned to one of three conditions (i.e., sing, aloud, or silent), which dictated how they would be instructed to study items. 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, one at a time, in one of three colors (red, yellow, or blue) and that they should ignore the color assignment of each word. Depending on the condition

to which they were assigned, participants were instructed either to (1) read all words silently, (2) read all words aloud, or (3) sing all words.

Study Phase. During the study phase, participants were presented with a series of 90 words, one at a time. Each trial began with a 500 ms fixation ("+"), followed by a 500 ms blank screen and then the word at center for 2000 ms. Depending on the condition to which they were assigned, participants either read each word silently, read each word aloud, or sang each word. Participants were supervised by an experimenter throughout the study phase. After all trials were complete, participants moved on 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 that were previously seen in the study phase and 90 of which were "new" foil words; all test words were presented in white font. 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. For the recollect/familiar/neither judgements, participants were instructed to respond with "recollect" to items that participants were able to visualize and for which they could recall subjective details about the associated encoding event, "Familiar" responses were to be given when participants recognized the item but were unable to recall subjective details. Finally, "neither" responses were

to be given when participants did not recognize the item, Responses for these judgements 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.

3.2.4 Statistical Approach

3.2.4.1 Signal Detection Analysis

The approach utilized in the present experiment was nearly identical to that described for Experiment 1, albeit with three exceptions. First, the models described below included only a fixed effect for condition (sing, aloud, silent) rather than fixed effects for condition and group. Second, the random effects structure of the models differed from Experiment 1 insofar as random slopes corresponding to participant-level variability across conditions were removed; because condition was manipulated between-subject in the present experiment, the inclusion of these model terms was no longer justified by the design of the experiment (see, e.g., Barr et al., 2013; Gelman & Hill, 2006). Finally, whereas Experiment 1 recorded only old/new responses as a dependent variable, the present experiment used confidence ratings and recollect/familiar/neither judgements. Accordingly, each dependent measure for the present experiment (i.e., confidence, recollection, and familiarity) was analyzed using a separate probit model. Further details about the parameterization of each model are discussed below. Finally, it is important to note that the design of this experiment intrinsically produces separate false alarm rates across conditions; thus, estimates of *C* produced by models reported hereafter could be meaningfully interpreted.

Priors and sampling procedures were identical to that described for the signal detection models reported in Experiment 1. While visual inspection of the chains and R-hat statistics indicated that the chains mixed well and that the models converged, effective sample size was

lower for estimates of *C* relative to comparable estimates reported for Experiment 1. Nonetheless, effective sample size was greater than 3000 for these estimates and greater than 7000 in all other cases.

3.2.4.2 Diffusion Models

The general approach used for diffusion models of Experiment 2 was similar to that described for Experiment 1, albeit with several exceptions. First, the confidence judgements collected in the present experiment could not be self corrected; thus, only a single response time was recorded for each trial and no trials were discarded on the basis of ambiguous decisions. Second, the design of the present experiment necessitated different fixed and random effects structures. Like the signal detection models described above, the diffusion models I implemented for Experiment 2 included only fixed effects for condition (as well as "old" status, for drift rate), rather than condition and group. Further, given that study condition was known to participants prior to the start of each trial in this experiment (i.e., because participants studied all items in the same manner), all parameters in the diffusion models were permitted to vary as a function of condition. With respect to random effects, random slopes corresponding to participant-level variability across conditions were removed as described for the signal detection analysis above. All other aspects of the modelling approach were identical to that described for Experiment 1. Once again, separate models of correctness and binarized old/new responses produced identical inferences, but the latter converged poorly; as such, I again report only the results for the simpler models.

3.2.4.3 Analyses of Serial Position

The approach taken for the serial position analyses of Experiment 2 differed from that reported for Experiment 1 only insofar as the linear effects structure and dependent variable of

the models was modified to match the probit models of binarized confidence ratings described above. Model parametrization, priors, and sampling procedures were otherwise identical to those described for the serial position analyses of Experiment 1.

3.3 Results and Discussion

3.3.1 Signal Detection Analysis

Table 3.1 shows means and standard deviations for all primary dependent measures as a function of condition. Table 3.2 shows means and standard deviations for confidence ratings and response times as a function of condition and item type.

3.3.1.1 Confidence Ratings

For my analysis of confidence ratings, responses were first binarized such that ratings greater than 3 indicated an "old" response.[14] I then applied a multilevel probit regression model to the binarized responses.

Sensitivity. Figure 3.1 depicts estimates for sensitivity across conditions and contrasts between conditions. As shown in Figure 3.1, a credible production effect was observed for both singing and reading aloud. However, no singing superiority effect emerged, and a non-credible numerical trend favored higher sensitivity for the aloud relative to the sing condition, difference = -0.15 ($HDI_{95\%}$ = -0.36 – 0.05). Interestingly, the production effects observed for both singing and reading aloud were comparable in magnitude to those observed for Experiment 1. 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 further discussed below in my analyses of recollection.

[14] I recorded confidence ratings because it had been my intention to analyze these data using a multilevel ordinal regression model. However, since conducting these studies, I have become aware that this is not yet possible in the manner I 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, I have instead adopted a more traditional approach.

Regarding random effects for sensitivity, the participant-level random intercept was informative (estimate = 0.41, $HDI_{95\%}$ = 0.34 – 0.48), as were item-level random slopes corresponding to the effect of condition (all estimates > 0.33) and item-level correlations between conditions (estimates > 0.51).

Table 3.1

Mean Proportion and Standard Deviation of the Mean for Hit Rates, Corresponding False Alarm Rates, Sensitivity (d'), and Response Bias (C) as a Function of Condition

Condition	Confidence				Recollection				Familiarity			
	Hits	FAs	*d'*	*C*	Hits	FAs	*d'*	*C*	Hits	FAs	*d'*	*C*
Sing	.59	.22	1.16	.31	.37	.13	1.19	.97	.56	.37	.62	-.03
	(.16)	(.19)	(.49)	(.63)	(.24)	(.24)	(.56)	(.98)	(.26)	(.33)	(.69)	(1.19)
Aloud	.65	.24	1.31	.21	.47	.15	1.26	.77	.61	.39	.75	-.09
	(.15)	(.20)	(.52)	(.60)	(.22)	(.21)	(.70)	(.86)	(.26)	(.30)	(.42)	(1.07)
Silent	.58	.35	.67	.09	.37	.19	.75	.73	.53	.41	.33	.03
	(.17)	(.20)	(.37)	(.61)	(.22)	(.22)	(.48)	(.86)	(.25)	(.26)	(.96)	(.81)

Note. Familiarity scores were computed using the independence remember/know procedure (Yonelinas & Jacoby, 1995; see below for further discussion). Sensitivity (*d'*) and response bias (*C*) were calculated conventionally (rather than estimated via Bayesian probit regression).

Table 3.2

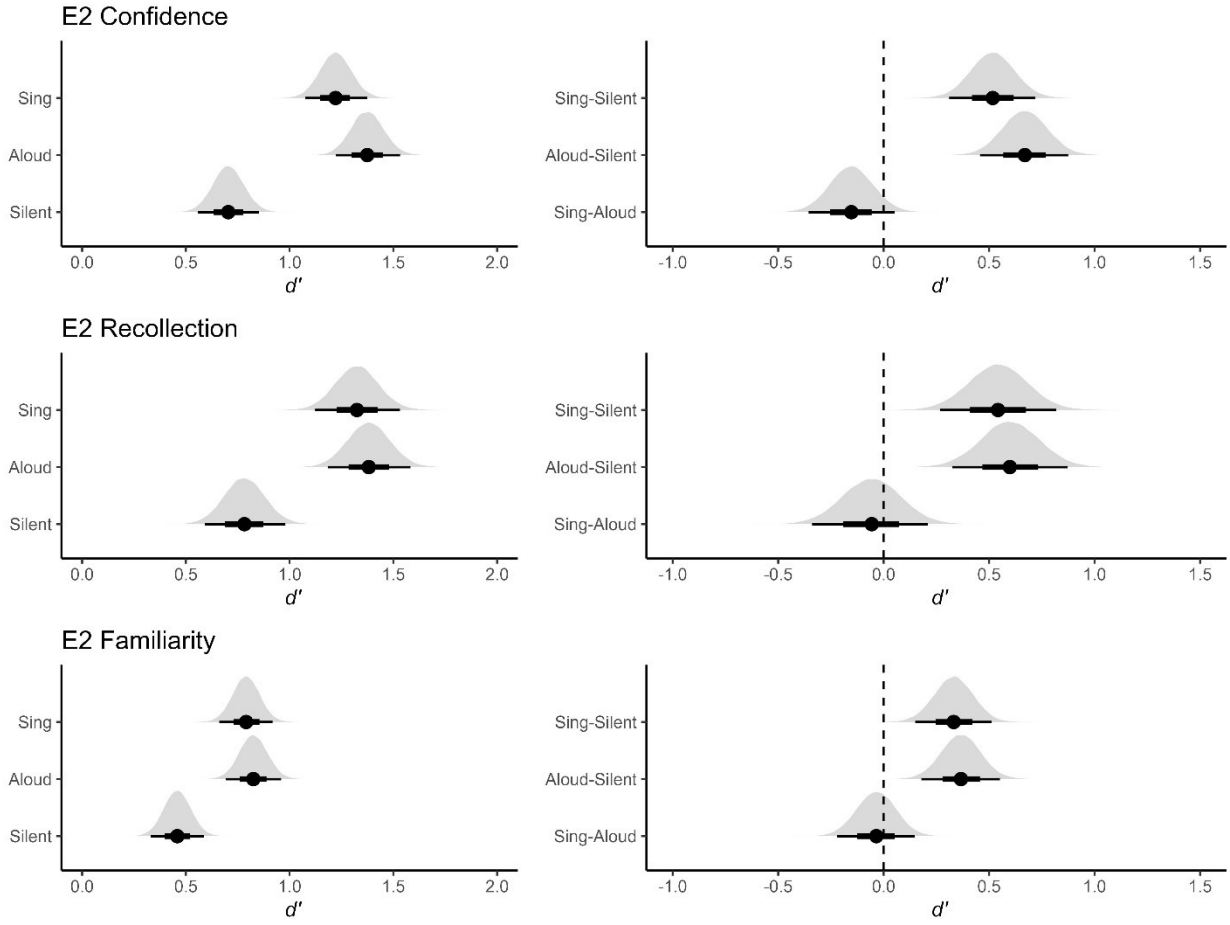
Mean and Standard Deviation of the Mean for Confidence Ratings and Response Times (RTs) in Seconds as a Function of Condition and Item Type

Condition	Item Type	RT	Confidence
Sing	Old	2.32 (1.18)	3.87 (2.19)
	New	2.21 (1.20)	2.31 (2.88)
Aloud	Old	2.32 (1.22)	4.21 (2.07)
	New	2.34 (1.23)	2.38 (2.10)
Silent	Old	2.31 (1.21)	3.84 (2.23)
	New	2.27 (1.23)	2.80 (2.03)

Note. Descriptive statistics for RTs were calculated after the removal of outlier trials, as outlined above.

Figure 3.1

Posterior Estimates for Sensitivity (d') as a Function of Condition (Left Column) and Contrasts Between Conditions (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.

Response Bias. Figure 3.2 depicts estimates for response bias across conditions and contrasts between conditions. No credible differences in response bias between conditions emerged, although a numerical trend favored more conservative response bias for the sing condition relative to the silent condition (difference = 0.23, $HDI_{95\%}$ = -0.04 – 0.49). Although

this pattern differs from that observed in Experiment 1, these results are consistent with earlier literature that has failed to observe production-related differences in response bias between-subject (e.g., Fawcett & Ozubko, 2016); this may suggest that experimental design modulates production effects on response bias. However, the trend towards an effect for the sing condition might also suggest that the effect is simply smaller in between- relative to within-subject designs, much like the production effect on sensitivity; thus, reliably detecting such an effect might require larger samples.

For response bias, the participant-level random intercept was informative (estimate = 0.61, $HDI_{95\%}$ = 0.54 – 0.71). Item-level random slopes corresponding to the effect of condition were also informative (estimates > 0.26), as were item-level correlations between conditions (estimates > 0.71).

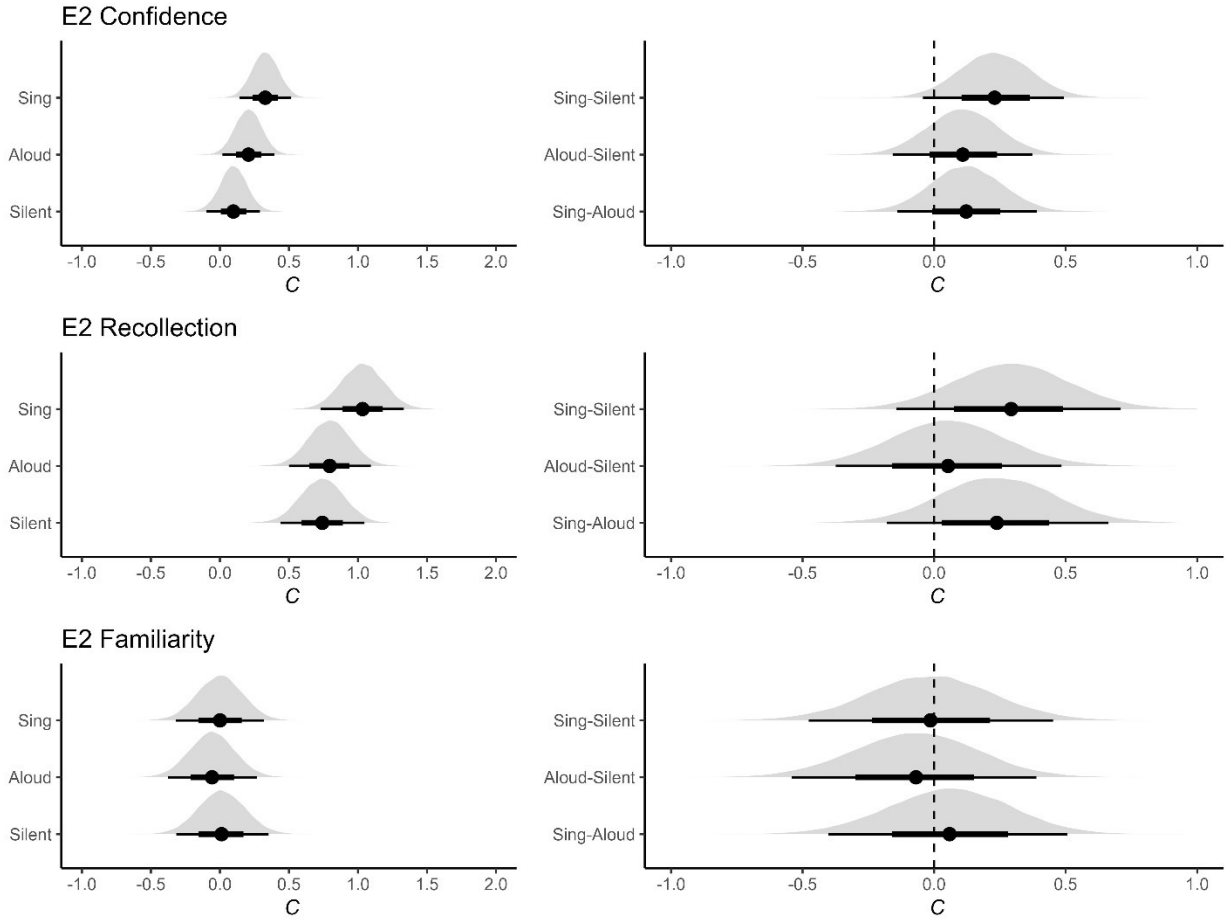
3.3.1.2 Recollection

Having evaluated the between-subject production effect for singing in standard recognition, I next applied a comparable multilevel probit model to analyze "recollect" responses. 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. Figure 3.1 depicts estimates for recollection sensitivity across conditions and contrasts between conditions. As shown in Figure 3.1, production effects were observed for either modality. As in my analysis of confidence ratings, there was no evidence for a credible singing superiority effect (difference = -0.06, $HDI_{95\%}$ = -0.34 – 0.21). Interestingly, however, the presence of a production effect for recollection in a between-subject design fails to replicate the findings of Fawcett and Ozubko (2016).

Figure 3.2

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

Given that the emergence of a between-subject production effect for recollection coincides with an unusually large production effect on confidence ratings for this design, it may be the case that some unknown methodological aspect of this experiment caused the recollective component to re-appear. To elaborate, Fawcett and Ozubko (2016) speculated that the between-

subject production effect is smaller than its within-subject counterpart because it is driven solely by processes related familiarity, rather than both recollection and familiarity; thus, if the recollective component of the benefit were to emerge between-subjects, the size of the production effect might be comparable across designs. Although no candidate moderators are apparent, the present experiment 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 in that investigation, 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, I cannot satisfactorily account for my observation of between-subject production effects on recollection; further research is necessary to elucidate the mechanisms that might have driven this pattern of results.

Much like my analysis of confidence ratings, the participant-level random intercept for the model of recollection was informative (estimate = 0.54, $HDI_{95\%}$ = 0.45 – 0.64), indicating baseline variability in recollection sensitivity across participants. Item-level random slopes reflecting variability in the effect of condition were also informative (estimates > 0.37), as were item-level correlations (estimates > 0.63).

Response Bias. Figure 3.2 depicts estimates for recollection response bias across conditions and contrasts between conditions. Consistent with the pattern observed for confidence

ratings, no credible differences in response bias emerged across conditions, although numerical trends favored more conservative bias in the sing condition relative to either the silent (difference = 0.29, $HDI_{95\%}$ = -0.15 – 0.71) or aloud condition (difference = 0.24, $HDI_{95\%}$ = -0.18 – 0.66). Consistent with earlier research (e.g., Fawcett & Ozubko, 2016), participants were considerably more conservative in making "recollect" responses relative to "old" confidence ratings; this is generally congruent with the notion that recollection represents a detailed reexperience of the encoding event and thereby requires a stronger mnemonic signal (e.g., Yonelinas, 2002).

For this parameter, the random intercept indicated that participants exhibited considerable baseline variability (estimate = 0.99, $HDI_{95\%}$ = 0.86 – 1.13). For item-level effects, all random slopes (estimates > 0.24) and correlations between conditions (estimates > 0.59) were informative.

3.3.1.3 Familiarity

Finally, I analyzed familiarity, which is often viewed as a nonspecific feeling of fluency or knowing 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 the present analyses, however, I 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 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. Figure 3.1 depicts estimates for familiarity sensitivity across conditions and contrasts between conditions. As depicted in Figure 3.1, analysis of the familiarity responses followed the same pattern observed for the other dependent measures, with credible production effects for either modality. Thus, it appears that the production effect for singing is driven at least in part by processes related to familiarity in between-subject designs. Once again, however, I observed no evidence for a singing superiority effect (difference = -0.03, $HDI_{95\%}$ = -0.22 – 0.15).

Once again, the random intercept corresponding to baseline variability in sensitivity across participants was informative (estimate = 0.30, $HDI_{95\%}$ = 0.23 – 0.38). The same was true for item-level random slopes corresponding to the effect of condition, although estimates for this parameter were numerically lower than in previous models (estimates > 0.12), suggesting less item-level variability in familiarity. Interestingly, correlations between conditions were not informative for this model and failed to reach credibility.

Response Bias. Figure 3.2 depicts estimates for familiarity bias across conditions and contrasts between conditions. Echoing trends observed in previous models for Experiment 2, no credible differences in response bias between conditions emerged.

Similar to my analysis of recollection bias, the participant-level random intercept indicated substantial variability in bias across participants (estimate = 1.06, $HDI_{95\%}$ = 0.92 – 1.23). Item-level random slopes for condition were less informative but credible nonetheless (estimates > 0.28), and correlations between conditions were positive and informative (estimates > 0.61).

3.3.2 Diffusion Models

I applied a multilevel diffusion model to accuracy coded from binarized confidence ratings. As described in Section 2.3.2, a single value of drift rate was computed for each condition. For Experiment 2, credible production effects on drift rate emerged for both reading aloud (difference = 0.38, $HDI_{95\%}$ = 0.24 – 0.53) and singing (difference = 0.31, $HDI_{95\%}$ = 0.17 – 0.45), indicating that participants accumulated evidence towards correct responses at a greater rate for produced relative to unproduced items. However, no singing superiority effect emerged for this parameter, with a numerical trend instead favoring lower drift rate for singing relative to reading aloud (difference = -0.07, $HDI_{95\%}$ = -0.21 – 0.07). Consistent with my diffusion models of Experiment 1, these results provide further evidence that the production effect can be observed in drift rate and extends this finding to a between-subject design.

With respect to other diffusion parameters, the design of Experiment 2 permitted separate estimates of boundary separation and nondecision time to be computed for each condition. Thus, I was able to explore production-related differences in these parameters in greater depth relative to my diffusion models of Experiment 1. The model revealed a strong but non-credible numerical

trend favoring higher boundary separation in the aloud relative to silent condition (difference = 0.17, $HDI_{95\%}$ = -0.03 – 0.38), and a similar but less pronounced trend emerged when comparing the sing and silent conditions (difference = 0.09, $HDI_{95\%}$ = -0.12 – 0.30). As with drift rate, there was a non-credible trend towards lower boundary separation in the sing relative to aloud condition (difference = -0.08, $HDI_{95\%}$ = -0.30 – 0.13). Although neither comparison with the silent condition reached credibility, these results are generally congruent with the notion that participant require additional evidence to make decisions for produced relative to unproduced items. From one perspective, this finding could be interpreted as consistent with the use of a heuristic strategy at test, wherein participants might think more carefully about items for which distinctive information can be retrieved relative to items for which retrieval fails. On the other hand, one might instead expect the inverse pattern: Once information about having produced an item is retrieved, participants might be expected to make a decision immediately, whereas participants might search for longer or employ additional strategies if retrieval of distinctive information fails (see Ozubko et al., 2020).

Finally, I observed non-credible trends favoring lower nondecision time relative to the silent condition for both the aloud (difference = -0.05, $HDI_{95\%}$ = -0.13 – 0.02) and sing conditions (difference = -0.05, $HDI_{95\%}$ = -0.13 – 0.03), but the difference between the two production conditions was centered on zero (difference = 0.00, $HDI_{95\%}$ = -0.07 – 0.08). While this pattern of results may appear at odds with the estimates for boundary separation discussed above, lower nondecision time does not necessarily equate to faster overall responses; rather, these findings suggest that for produced items, participants spent less time encoding the stimuli prior to initiating the decision processes—thereby still allowing for longer decision processes and overall response times for produced items. To contextualize these results in light of a

distinctiveness heuristic, it could be that decision processes are initiated faster when distinctive information can be retrieved successfully. However, accounts that assume increased encoding strength for produced items might also predict such a pattern: Stronger memory traces might be more easily retrieved, leading to faster process initiation.

3.3.3 Analyses of Serial Position

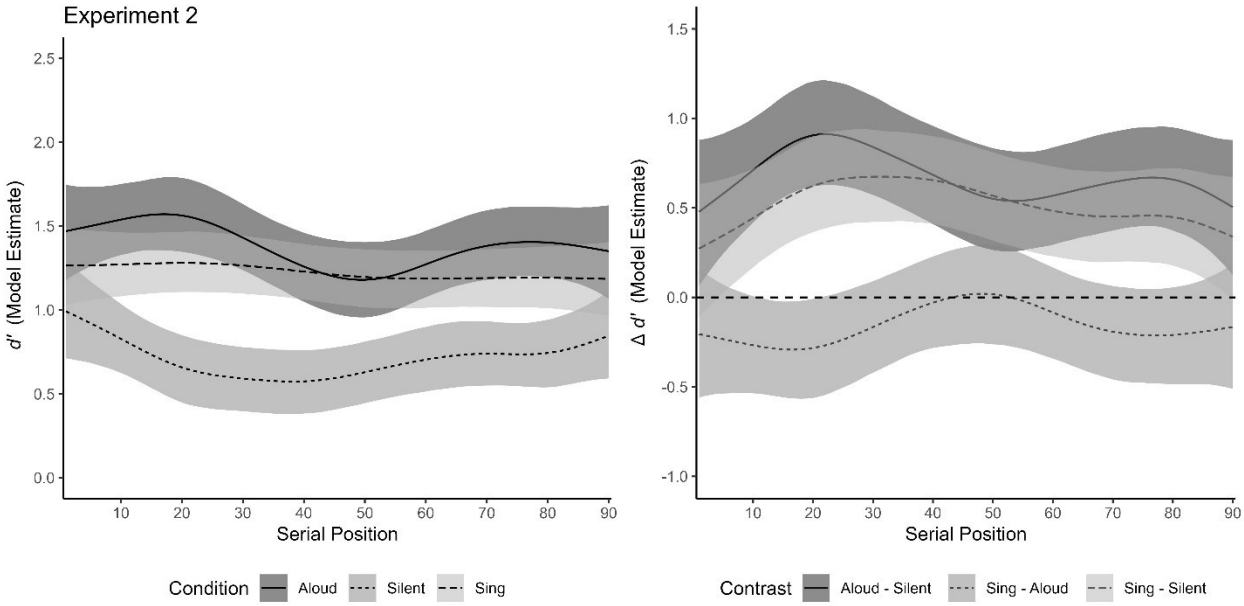
Figure 3.3 depicts conditional nonlinear model estimates for *d*' across serial positions and contrasts between conditions across serial positions for Experiment 2.

For the linear model, no credible evidence for an interaction between production and serial position emerged. In this case, model comparison provided evidence favoring the nonlinear model over the linear model (ΔELPD = -14.6, ΔSE = 4.9). Because participant-level random slopes for serial position were not justified by the design of this experiment, no additional linear models were fit. Comparing the nonlinear model to probit models of confidence ratings excluding serial position effects provided evidence favoring the inclusion of such effects (ΔELPD = -21.0, ΔSE = 6.6).

As shown in Figure 3.3, a nonlinear trend emerged for the aloud condition such that the production effect for either modality increased in size early in the study list (peaking at roughly position 20) and subsequently decreased. In this model, the production effect was credible for reading aloud at all serial positions. For singing, the effect was credible from positions 5 through 89 and marginal at all other positions. For this experiment, planned contrasts provided no credible evidence favoring a larger production effect for late relative to early positions in the aloud condition (difference = 0.02, $HDI_{95\%}$ = -0.38 – 0.41), and only a weak trend emerged in the sing condition (difference = 0.06, $HDI_{95\%}$ = -0.33 – 0.43).

Figure 3.3

Nonlinear Model Estimates for d' as a Function of Condition and Serial Position (Left Panel) and Nonlinear Model Estimates for the Production Effect (Sing/Aloud - Silent) and the Singing Superiority Effect (Sing - Aloud) as Function of Serial Position (Right Panel)



Note. The shaded region surrounding each curve depicts the 95% quantile interval of the estimate.

Unlike Experiment 1, I found no evidence for practice effects in any modality for Experiment 2. Otherwise, these findings are only weakly consistent with the meta-analytic evidence reported in Fawcett et al. (2023): Although the production effect was larger for late relative to early positions for either modality, the differences were very small and did not approach credibility. Thus, like Gionet et al. (in press), my analysis of Experiment 2 failed to provide strong evidence for the interaction predicted by the RFM in pure-list recognition paradigms. Nonetheless, my findings disagree with Gionet et al. insofar as the production effect

does interact with serial position to some degree, but not necessarily in the predicted manner: The size of the production effect for either modality varied systematically across serial positions, peaking early in the list. However, the theoretical implications of this trend are not obvious and further research is needed to elucidate the specific nature of these findings. Given that this unusual trend was not captured by the linear model, future investigations of serial position and production might benefit from approaches capable of capturing nonlinear effects. Finally, with respect to practice effects in Experiment 2, I found no evidence suggesting that performance in any condition improved as serial position increased.

Chapter 4: Meta-analysis of the Singing Superiority Effect

4.1 Overview

The purpose of this investigation was to address disagreement amongst findings regarding the singing superiority effect by extending the meta-analysis of the effect reported in Whitridge (2022). As discussed in Section 1.3 of the present thesis, the results of investigations into the production effect for singing vary widely: While Quinlan and Taylor (2013, 2019) reported large singing superiority effects, both Hassall et al. (2016) and Whitridge (2022) found a production effect for singing that was similar in magnitude to that observed for reading aloud. Accordingly, Whitridge (2022) aggregated all known studies that compared performance for reading aloud and singing in production paradigms in order evaluate evidence for the effect as a whole. The author found that the aggregate singing superiority effect was significant, albeit small and with evidence of heterogeneity: The meta-analytic model revealed that the distribution of effects expected from a typical study ranged from large singing superiority effects (Hedge's g = 1.05) to moderate effects favoring a benefit for reading aloud over singing (Hedge's g = -0.39).

The purpose of updating the meta-analytic model reported in Whitridge (2022) was twofold. First, Experiment 1 in the present thesis contributed two new independent effects that could be added to the model (i.e., the matched and unmatched groups), and I became aware of an additional single-effect study that met my inclusion criteria (Zhang, 2024); increasing the number of effect sizes present in the model will allow for a better estimation of the aggregate effect and between-study heterogeneity. Second, whereas Whitridge (2022) conducted a Frequentist meta-analysis, I adopted a Bayesian approach in the present investigation, the advantages of which I discuss below. For this investigation, I predicted a pattern of results akin to Whitridge (2022): I expected to observe evidence favoring an aggregate singing superiority effect, but I also expected to observe evidence for substantial heterogeneity across studies.

4.2 Method

4.2.1 Search and Coding

As in Whitridge (2022), the present meta-analysis made use of a recent search already conducted for the production effect literature (Fawcett et al., 2023). 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 a forward and backward snowball search of articles included in their search. For the present investigation, the search conducted by Fawcett et al. (2023) was combined with forward and backward snowball searches of each study that previously contributed effect sizes to the meta-analysis reported in Whitridge (2022). Additionally, I conducted forward and backward snowball searches of two potentially relevant articles that manipulated singing in language acquisition paradigms (i.e., Baills et al., 2021; Ludke et al., 2014) to determine whether any studies in this area met the inclusion criteria.

The combined search efforts did not identify any additional articles that met the inclusion criteria beyond those previously included in Whitridge (2022). Although some articles within the language acquisition literature included manipulations of singing versus reading aloud, these studies were excluded largely due to immense procedural variation relative to typical production paradigms. However, I became aware of an eligible unpublished investigation of the production effect for singing (Zhang, 2024) via correspondence with the author, for which I was able to obtain raw data. Accordingly, the present analyses included five studies based on their reporting a within-subject production manipulation including both sing and aloud conditions: Quinlan and Taylor (2013, 2019), Hassall et al. (2016), Zhang (2024) and Whitridge (2022). Including the effects reported in the present study, my sample consisted of 14 independent effect sizes reported across five studies. Means, standard deviations, sample sizes and correlations between the sing and aloud conditions were recorded for each included experiment.

4.2.2 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 the primary dependent measure across my own experiments has been sensitivity (rather than raw or corrected hits), I 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), I imputed this parameter using the other data available to me.[15]

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. I 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).[16] 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).

I modeled the data using two approaches, first including the singing superiority effect as the effect of interest. Subsequently, I fit separate models including the production effect observed in the singing and read aloud conditions with production modality as a moderator. In either case, the general parameterization I used was analogous to a Frequentist random effects meta-analysis, such that the models estimated the size of the aggregate effect across studies weighted by sampling variance and included random effects corresponding to the experiment from which each effect was derived, thus assuming variability in effect sizes across studies; given heterogeneity across samples, sites, and methodologies used in investigations of the production effect for singing, I believe this assumption to be justified (for further discussion, see, e.g., Borenstein et al., 2010). However, some aspects of the parameterization differed across my approaches.

[15] In this case, I imputed the missing within-subject correlation by taking the average of all within-subject correlations that were available to me (see, e.g., Furukawa et al., 2006).
[16] 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).

For models of the singing superiority effect, I computed the dependent measure as the raw mean difference in *d'* (calculated conventionally; see, e.g., Stanislaw & Todorov, 1999) between the sing and aloud conditions. Models fit using this approach included a random intercept that permitted baseline variability in the size of the singing superiority effect across studies. With this parameterization in mind, the models computed an intercept corresponding to the estimated aggregate singing superiority effect across studies. Additionally, the models produced a random intercept corresponding to between-study heterogeneity, akin to Tau in Frequentist models (see, e.g., Harrer et al., 2021). For models of the production effect, I computed the dependent measure as the raw mean difference in *d*' between produced (sing/aloud) and unproduced conditions. These models included a fixed effect for production modality, a random intercept to quantify between-study heterogeneity, and a random slope that assumed variability in the impact of production modality across samples. Because the models of the production effect used a common comparison condition for each effect (i.e., silent), my approach allowed for the estimation of separate dependent effects for each sample.

I also applied uninformative, mildly regularizing priors to the meta-analytic models. For models of the singing superiority effect, these priors reflected my 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 my *a priori* theoretical belief that a slightly more elaborate type of vocalization should not be vastly superior to an already large benefit; nonetheless, these priors did allow for the possibility of very large singing superiority effects in individual studies. For models of the production effect,

my priors were specified in accordance with my belief that the raw mean difference in *d'* between either production condition and the silent condition should reasonably fall between -1 and 1 in a typical study, with effects in individual studies permitted to range from -2 to 2. These priors were calibrated with respect to effects reported in previous investigations of the production effect for singing (i.e., Quinlan & Taylor, 2013, 2019; Hassall et al., 2016; Whitridge, 2022).

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 1.01 in all cases, indicating that all models converged (Gelman & Hill, 2006; Kruschke, 2010). Further, the effective sample size was greater than 70000 for all estimates.

4.3 Results and Discussion

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

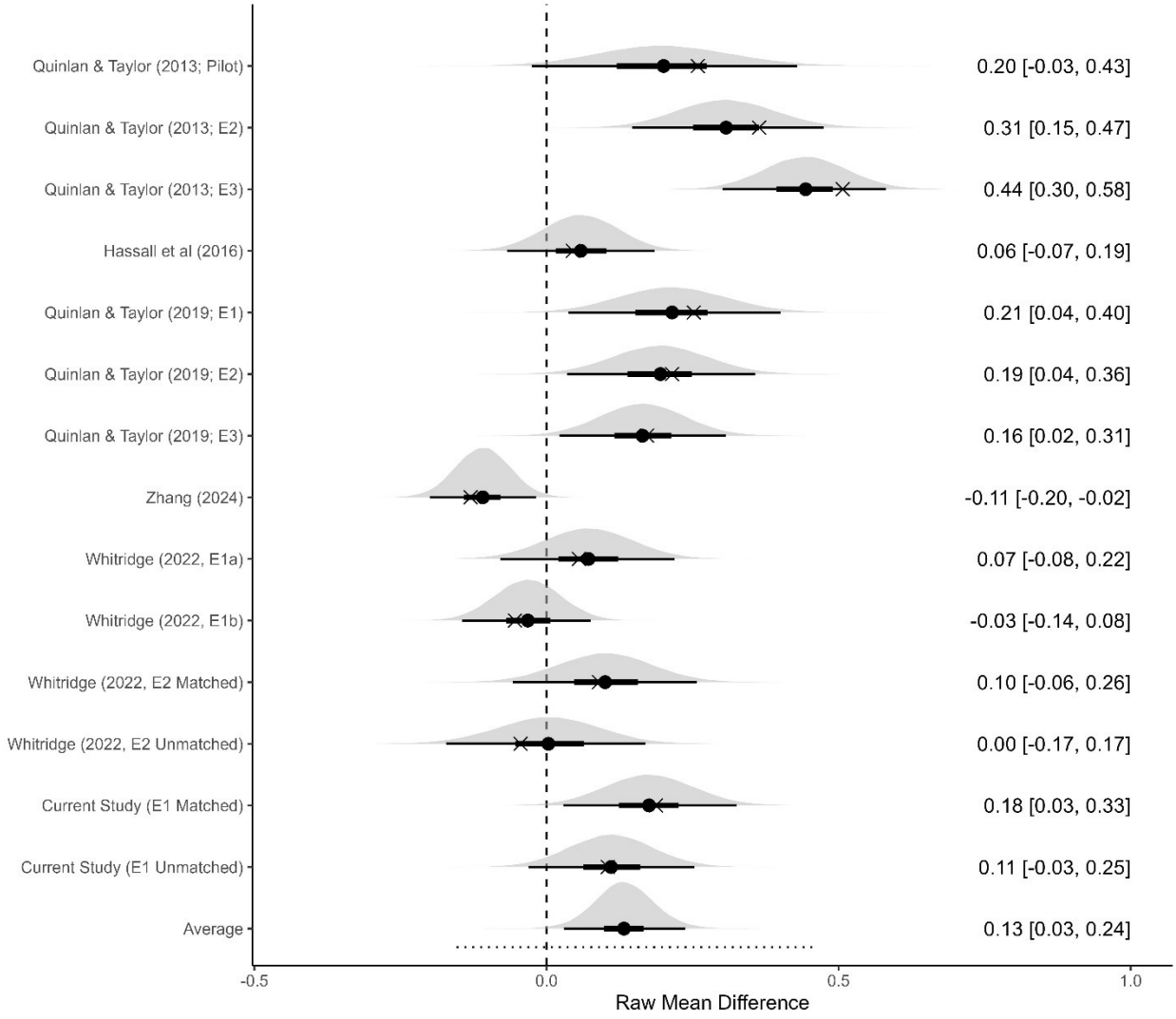
4.3.1 Models of the Singing Superiority Effect

Figure 4.1 shows a forest plot of the meta-analytic model of the singing superiority effect. As depicted in Figure 4.1, the aggregate singing superiority effect was credible, with the difference between the sing and aloud conditions estimated at 0.13 ($HDI_{95\%}$ = 0.03 – 0.24). 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 revealed substantial heterogeneity across reported effects, with prediction intervals ranging from -0.14 to 0.47; this implies that some studies show unsupportive effects (i.e., aloud > sing),

whereas others show effects that are quite large. The random intercept permitting baseline variability in the size of the singing superiority effect was also informative (estimate = 0.17, $HDI_{95\%}$ = 0.10 – 0.27), further supporting the notion that there is substantial between-study heterogeneity in the size of the singing superiority effect. 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). Like the meta-analysis reported in Whitridge (2022), this model suggests that the singing superiority effect – if truly reliable – has likely been overestimated.

Figure 4.1

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.

Given that Experiment 1 hinted at the possibility that color matching might play an important role in facilitating the singing superiority effect (see also, Whitridge, 2022; Experiment 3), I fit an exploratory meta-analytic model 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 the model now included a categorical fixed effect, I removed the model intercept and instead computed slopes corresponding to the aggregate singing superiority effects for color matched and unmatched studies, respectively (akin to a subgroup analysis; see, e.g., Borenstein & Higgins, 2013). Here, the aggregate singing superiority effect was credible when color matching was present, with the difference between the sing and aloud conditions estimated at 0.22 ($HDI_{95\%}$ = 0.11 – 0.33; $PI_{95\%}$ = -0.00 – 0.49). However, the effect was not credible in the absence of this procedure, estimated at 0.00 ($HDI_{95\%}$ = -0.11 – 0.14; $PI_{95\%}$ = -0.22 – 0.29). Interestingly, the random intercept corresponding to between-study heterogeneity was numerically smaller and less informative relative to that observed in the previous model (estimate = 0.12, $HDI_{95\%}$ = 0.06 – 0.21), suggesting that there was less heterogeneity across studies after accounting for the use of color matching. Consistent with the trend observed in Experiment 1, these results suggest that the singing superiority effect might emerge only when color matching is used at test.

4.3.2 Models of the Production Effect

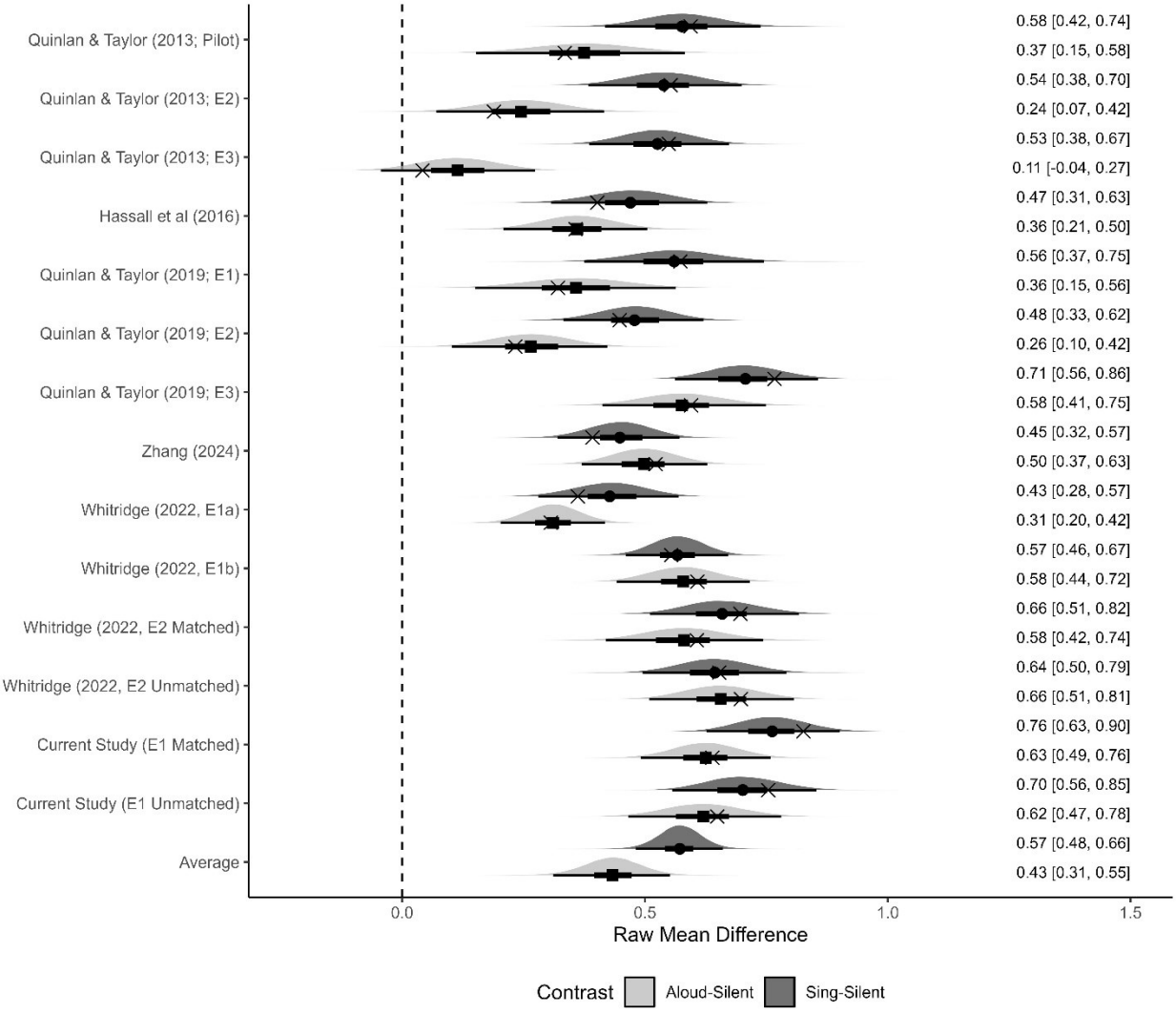
Figure 4.2 depicts a meta-analytic model of the production effect. As shown in Figure 4.2, the aggregate production effects for both aloud and sing conditions were credible, with the differences between the sing/aloud and silent conditions respectively estimated at 0.57 ($HDI_{95\%}$ = 0.48 – 0.66; $PI_{95\%}$ = 0.35 – 0.82) and 0.43 ($HDI_{95\%}$ = 0.31 – 0.55; $PI_{95\%}$ = 0.07 – 0.74). The

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

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

Figure 4.2

Forest Plot Depicting Raw Mean Differences in d' (Aloud-Silent and Sing-Silent) for a Meta-Analytic Model of the Production Effect



Note. Light-coloured polygons and square points represent the difference in *d'* between the aloud and silent conditions, whereas dark-coloured 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.

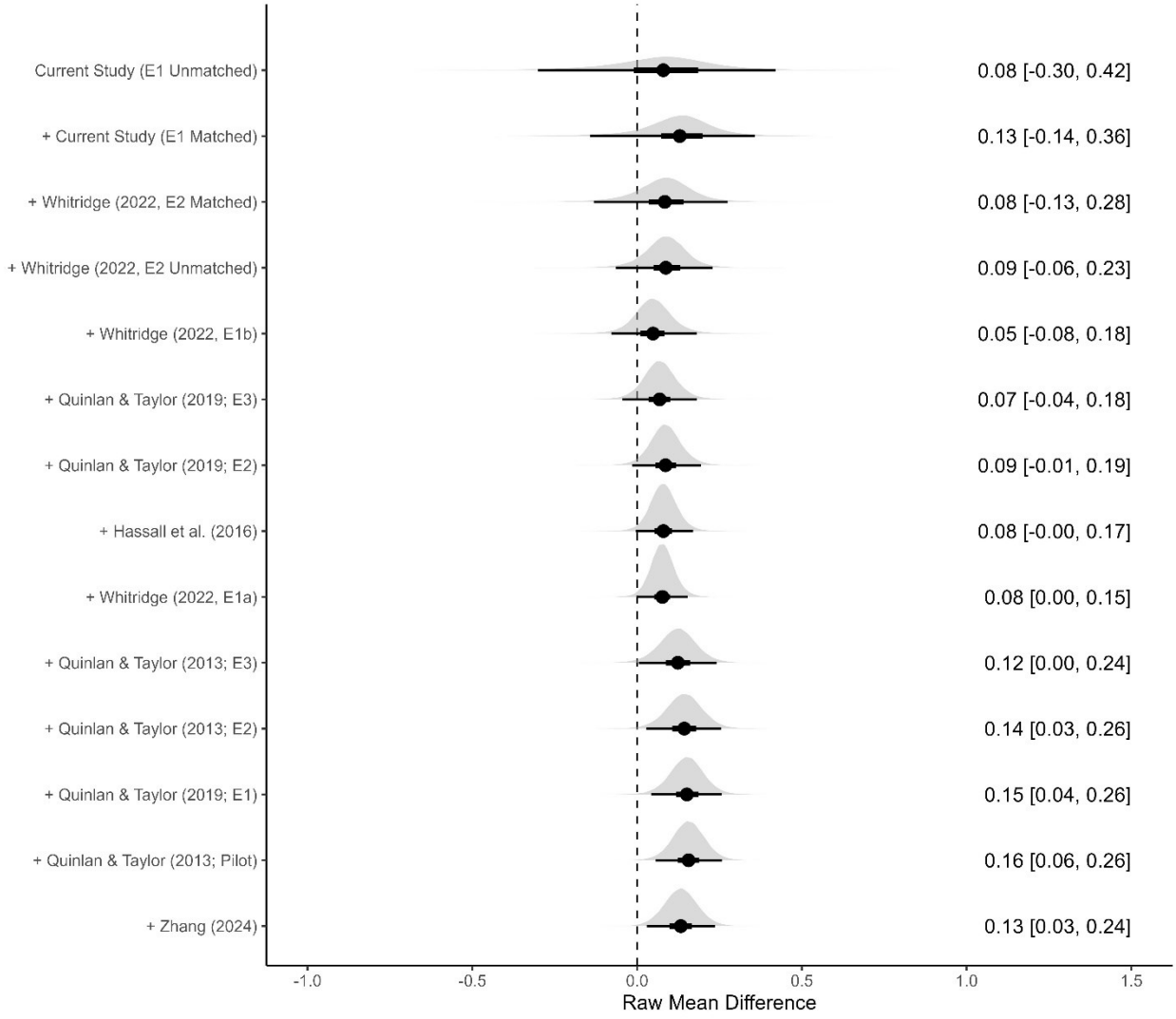
4.3.3 Analysis of Publication Bias

Finally, to evaluate publication bias, I 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 my other meta-analytic models of the singing superiority effect. These models were parameterized similarly to conventional calculations of Egger's regression test, albeit inclusive of random effects; the random effect structure of these models and the corresponding assumptions were identical to that reported for the models 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 singing superiority effect 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 (reflecting the effect size for a study with an average sample size) was estimated at 0.14 ($HDI_{95\%}$ = 0.04 – 0.24) and the slope for sample size narrowly failed to reach credibility (estimate = 0.08, $HDI_{95\%}$ = -0.03 – 0.19). For the model including standard error, the intercept (reflecting the effect size for a study with an average standard error) was estimated at 0.13 ($HDI_{95\%}$ = 0.03 – 0.24) and the slope for standard error was not credible (estimate = -0.04, $HDI_{95\%}$ = -0.15 – 0.08). Thus, these analyses failed to provide convincing evidence for publication bias. For either model, the estimates corresponding to the random intercept were informative and comparable in size to that of my first meta-analytic model.

To further evaluate publication bias, I fit a cumulative meta-analysis of the singing superiority effect, 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, I began by fitting a model of the singing superiority effect that was parameterized identically to my first meta-analytic model reported above, albeit including only the largest study. Subsequently, I added the next largest study and re-fit the model; this procedure was repeated until all studies had been included. For all iterative models, priors, sampling procedures and random effects structures were identical to my other meta-analyses of the singing superiority effect. As shown in Figure 4.3, the aggregate singing superiority effect 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 singing superiority effect 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).

Figure 4.3

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



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).

Chapter 5: General Discussion

5.1 Overview of Results

The central purpose of the present thesis was to extend the evaluation of the production effect for singing conducted by Whitridge (2022). Previous investigations within this area have produced conflicting findings, with two studies reporting large singing superiority effects (Quinlan & Taylor, 2013, 2019) and two studies observing a production effect for singing that was comparable in magnitude to that for reading aloud (Hassall et al., 2016; Whitridge, 2022). Despite these discrepant findings, the singing superiority effect has been accepted as evidence for the distinctiveness account of the production effect (e.g., Forrin & MacLeod, 2018; Mama & Icht, 2016): Proponents argue that the relative superiority of singing arises on the basis of additional distinctive features encoded with the production trace (Quinlan & Taylor, 2013, 2019), which validates the sensorimotor scaling corollary of the distinctiveness account (Fawcett et al., 2012; Forrin et al., 2012). While Whitridge (2022) posed a strong initial challenge to this claim, the present investigations explored two important points not yet addressed: (1) that the singing superiority effect might be driven in part by hidden moderators related to study design, and (2) that the production effect for singing should be robust in between-subject designs.

To address these points, I first modified the conceptual replications reported in Whitridge (2022) to replicate the methods of Quinlan and Taylor (2013; Experiment 3) as exactly as possible; this entailed the modification of items, the addition of familiarization and practice phases, and the inclusion of a group for whom test items were presented in the corresponding font colour from study. Although a credible singing superiority effect emerged in this experiment, it was confined to the color matched group and was smaller than those previously reported. Subsequently, in Experiment 2, I evaluated the production effect for singing using a

between-subject paradigm that conceptually replicated Quinlan and Taylor (2019; Experiment 4). Here, I detected a credible production effect for singing that was of similar magnitude to that for reading aloud; however, no credible singing superiority effect emerged. Finally, to address discrepancy amongst findings within the literature, I updated the meta-analysis of the singing superiority effect reported in Whitridge (2022) and conducted exploratory moderator analyses to evaluate the possibility of an interaction between singing and color matching. This investigation provided evidence for a credible aggregate singing superiority effect, but the effect was smaller than previously estimated (e.g., Quinlan & Taylor, 2013, 2019) and appeared to be driven by use of the color matching procedure—and inflated by small sample effects. Considered in aggregate, the results reported herein support three conclusions: first, that the size of the singing superiority effect has been overestimated by previous investigations; second, that the effect arises at least partially on the basis of factors related to study design, which limits the support this effect can provide for the sensorimotor scaling hypothesis; and finally, that the production effect for singing persists between-subjects, and is thereby not likely driven solely by relative distinctiveness. In the discussion that follows, I contextualize my findings with respect to theoretical frameworks of the production effect and the mnemonic utility of singing.

5.2 Implications for Scaling Distinctiveness

Earlier investigations of the singing superiority effect have explained the benefit with reference to the idea of scaling distinctiveness: Singing is thought to append additional distinctive features to the production trace in the form of sensorimotor information related to pitch or rhythm (Quinlan & Taylor, 2013, 2019). Although other studies have shown the production effect to decrease when distinctive features are removed in experimental paradigms (e.g., Forrin et al., 2012; Mama & Icht, 2016), evidence for the inverse of this pattern rests solely

upon the singing superiority effect, despite exhaustive efforts to increase the size of the production effect (e.g., Ozubko et al., 2020; Wakeham-Lewis et al., 2022). However, evidence that the singing superiority effect is in fact driven by scaling distinctiveness derives almost entirely from evidence that the effect is *not* driven by other mechanisms. Across their investigations, Quinlan and Taylor (2013, 2019) provided evidence that the relative superiority of singing did not arise due to increased intensity, bizarreness, or production time. The only prior experiment to directly address a distinctiveness-based explanation did so by manipulating production between-subjects, but some investigations have since provided evidence that discrepancies in the size of the production effect across designs are compatible with both distinctiveness- and strength-based frameworks (Jamieson et al., 2016). Given the lack of direct evidence for the sensorimotor scaling account of the singing superiority effect, then, I believe it is important to evaluate the findings of the present investigation with respect to the predictions that would be made by such an account.

For example, it can be inferred that the sensorimotor scaling hypothesis would predict the singing superiority effect to be both large and reliable. Given that previous investigations have observed substantial decrements to the size of the production effect when distinctive encoding processes are eliminated at study (e.g., a mean difference in hit rates between the produced and unproduced conditions of ~ 0.06 for writing compared to ~ 0.20 for reading aloud; Forrin et al., 2012), adding at least one additional encoding process via singing would be expected to produce a comparable advantage. Furthermore, although some degree of between-study heterogeneity is to be expected, the presence of additional distinctive features should be inherent to a given output modality. Relative advantages thought to be driven by the presence of additional encoding processes should thereby emerge reliably across investigations despite minor discrepancies in

methodology, which has been shown by previous work: Forrin et al. (2012) replicated the relative superiority of reading aloud over mouthing first reported by Conway and Gathercole (1987). Accordingly, the sensorimotor scaling hypothesis would also predict the singing superiority effect to emerge reliably because changes in site or methodology should not impact the number of distinctive encoding processes elicited by the modality.

Across my own investigations, however, the singing superiority effect appears neither large nor reliable. For example, the initial effects observed by Quinlan and Taylor (2013) were mean differences in sensitivity of ~ 0.36 and 0.51 for Experiment 2 and Experiment 3, respectively. Conversely, the sole singing superiority effect I observed was much smaller, with the mean difference between the sing and aloud conditions estimated at only 0.23. Further, my meta-analytic model estimated a small aggregate singing superiority effect of 0.13, which was only more pronounced in studies using color matching (difference = 0.22) and negligible in unmatched studies (difference = 0.00). 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., $N = 27 - N = 43$) 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). Taken together with the results of my cumulative meta-analytic model, which suggested that the aggregate benefit was driven largely by small studies that observed large effects, it appears likely that large singing superiority effects previously reported reflect inflated estimates.

It is difficult to reconcile the lower magnitude of the singing superiority effect observed herein with the sensorimotor scaling hypothesis. Interpreted at face value, the account would suggest that additive benefits operate on an all-or-nothing basis, wherein the magnitude of the effect depends solely upon the number of encoding processes elicited at study irrespective of what these processes are (see, e.g., Mama & Icht, 2016). If one rejects or modifies this account, however, a small singing superiority effect is perhaps better aligned with expected patterns: Because typical production effects deriving from reading aloud already entail very large benefits to sensitivity relative to silent reading (e.g., *MD* = ~0.78; Forrin et al., 2016), it seems unlikely that a slightly more elaborate form of vocalization should nearly double the size of the effect. Rather than an all-or-nothing basis, then, it could be that additional features associated with a common modality possess less discriminative utility relative to those deriving from distinct modalities. To elaborate, previous tests of the sensorimotor scaling hypothesis have observed large effects when manipulating the presence of distinctive features by altogether removing processes associated with a given modality (e.g., auditory or visual processing; Forrin et al., 2012; Conway & Gathercole, 1987; Mama & Icht, 2016). On the other hand, singing is thought to afford additional distinctive features that must be encoded via auditory processing (e.g., pitch or rhythm; Quinlan & Taylor, 2013, 2019), which is also elicited by reading aloud. Thus, any processing related to pitch or rhythm might simply encode additional auditory features at study, rather than representing the addition of a wholly distinct form of processing. To accommodate this explanation, however, one must adopt a more nuanced model of distinctiveness than those outlined by previous research (e.g., Forrin et al., 2012; Mama & Icht, 2016).

In addition to being small in magnitude, my investigations largely suggest that the singing superiority effect emerges only in the presence of the color matching procedure. This

finding raises questions about the theoretical mechanisms driving the singing superiority effect, given that this procedure may alter how participants engage with items at test. Although the design of the present investigations did not allow me to definitively ascertain the mechanism through which singing might interact with color matching, one potential explanation is that presenting items in their study colors – and thereby orienting participants to the condition under which items were studied – encourages different strategies to be employed across conditions. In typical production paradigms, participants are thought to scan their memories for distinctive information about having produced items to adjudicate between studied items and foils (e.g., Dodson & Schacter, 2001; MacLeod, 2010; see also, Fawcett & Ozubko, 2016). However, if a participant is aware that an item would have been studied silently, they may abstain from scanning their memory for information about having produced it. On the other hand, knowing that a test item was sung or read aloud might encourage participants to search for or reactivate very specific sensorimotor information about having produced the item using a particular modality. With respect to why an alternative type of distinctiveness heuristic might preferentially lead to a larger production effect for singing, it is possible that additional 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 production benefit to emerge, the production trace must be both discriminative and utilized to guide retrieval; if participants typically neglect additional features specific to singing, no singing superiority effect would be expected.

An analogous alternative explanation for such an interaction is that colour matching might help to reinstate study context at test. Wakeham-Lewis et al. (2022) suggested that in

production paradigms, participants might consciously reinstate the study phase production condition 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 specifically imagine singing the item. According to this *sensorimotor reinstatement hypothesis*, recreating the productive act in one's mind would preferentially benefit singing (or other elaborate modes of speaking, e.g., character voices; 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 my 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.

While distinctiveness- and context-based accounts provide plausible (albeit speculative) explanations for the interaction between the singing superiority effect and color matching, these accounts do not fit neatly into extant sensorimotor scaling models of distinctiveness (e.g., Fawcett et al., 2012; Forrin et al., 2012; Mama & Icht, 2016). As formulated by Mama and Icht (2016), the sensorimotor scaling hypothesis contends that the magnitude of the production effect depends on the number of unique encoding processes elicited at study. However, the present investigation highlights that simply adding unique encoding processes to the study phase is not sufficient to afford additional discriminative utility to the production trace: If processing specific to singing encodes additional sensorimotor features relative to reading aloud, it appears that these features are either unavailable in typical paradigms or that participants simply do not leverage the features unless prompted to do so. Accordingly, these findings pose a strong challenge to

current formulations of the sensorimotor scaling hypothesis, suggesting instead that the presence of additional distinct processes at study does not inherently impact the magnitude of the production effect.

5.3 Alternative Accounts of the Singing Superiority Effect

Given the failure of the sensorimotor scaling hypothesis to adequately explain the patterns of findings observed across this investigation and in Whitridge (2022), it might be the case that an alternative mechanism drives the singing superiority effect. Indeed, perhaps the simplest explanation for my difficulty in replicating the effect is that singing simply 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 occur 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 would not predict a relative advantage for singing.

However, the scaling model 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, singing might allow for a greater degree of variation in item representations across articulatory and auditory features. To elaborate, although both singing and speaking involve pitch, the two modalities differ in how they use

pitch. For example, singing involves highly accurate use of pitches that are typically organized in a scale, whereas pitch in speech intonation is less precise and organized (e.g., Zatorre & Baum, 2012); further differences can be found in the articulatory demands of each modality (e.g., Burrows, 1989). Thus, while both singing and speaking may encode a common set of auditory and articulatory features, those deriving from singing might contain additional sensorimotor information that could be leveraged to guide discrimination at test. This hypothesis fits well with neuroimaging studies of singing, which have observed substantial processing differences between singing and speaking (e.g., Geiser et al., 2008; Jeffries et al., 2003; Özdemir et al., 2006). 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.

If singing allows for qualitative variation in item representations across sensorimotor features, it is possible that certain stimulus dimensions might render items more conducive to the emergence of a singing superiority effect. Earlier, I discussed the notion that participants might be more likely to sing multisyllabic items with greater melodic variation because it is more natural to do so. However, exploratory analyses of the experiments reported herein and of data from Whitridge (2022) failed to observe any credible interaction between number of syllables and singing. Thus, the qualitative account outlined above does not receive any immediate support from the present investigations. However, it could be the case that additional syllables are not sufficient to meaningfully increase variability, or that participants do not typically sing with much variability unless explicitly prompted to do so. Accordingly, this hypothesis could be tested

using longer stimuli that permit greater variation in pitch (e.g., sentences) or instructing participants to sing items using particular melodies. At present, however, evaluating the viability of a qualitative account requires further investigation.

Nonetheless, even a highly modified formulation of the sensorimotor scaling hypothesis would struggle to account for past and current difficulties in replicating the singing superiority effect. Of the relevant studies published prior to Whitridge (2022), Hassall et al. (2016) is unique insofar as that investigation failed to detect a singing superiority effect despite using color matching at test. Those authors explained their failure to replicate the effect with reference to methodological differences, suggesting that the effect failed to 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 my own failures to observe a singing superiority effect. Here and in Whitridge (2022), my 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, my participants were supervised throughout the study phase and prompted to sing more effortfully if their singing faltered at any point. Because previous efforts did not go as far as to implement these safeguards, it seems unlikely that my 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, my findings instead suggest that this advantage is itself atypical and can emerge only when certain conditions are met.

An alternative means of explaining the apparent inconsistency of the singing superiority effect might be to suggest that singing-related expenditure of cognitive resources could obfuscate

the superiority of the modality. Whereas reading aloud necessitates only that participants process the word and speak it, singing elicits additional demands: Participants must improvise a melody for each word, which requires the simultaneous execution of several cognitive processes (e.g., performance monitoring and perceptual encoding; Pressing, 1988). Increases to cognitive load are known to impair performance on tests of recognition memory (e.g., Jones et al., 2012), and this pattern of findings has been extended directly to production paradigms (e.g., Mama et al., 2018). Further to this point, earlier research has demonstrated that diversion of attention related to performance anticipation can hinder performance in production paradigms (Forrin et al., 2019). Given that singing in front of an experimenter would likely be considered more embarrassing than simply reading items aloud (see, e.g., Hofmann et al., 2006) and that the complexity of the former might require additional preparatory processing (Beaty, 2015), it is possible that attentional diversions related to performance anticipation could be accentuated for this modality. Thus, the additional cognitive demands elicited by singing (anticipatory or otherwise) could diminish benefits arising from the presence of additional sensorimotor features.

However, previous studies of production and singing appear to argue against these points (Quinlan & Taylor, 2019; Hassall et al., 2016). Because the cognitive processes elicited by musical improvisation become automatized with practice (Pressing, 1988), the cognitive load associated with this modality is diminished for experienced musicians (Beaty, 2015). If an attentional mechanism hinders benefits arising from singing, then, this population might be expected to show a more pronounced singing superiority effect. Contrary to this hypothesis, however, Quinlan and Taylor (2019) observed a singing superiority effect of typical size in a sample of singers who had at least one year of experience. With respect to explanations related to performance anticipation, neuroimaging work by Hassall et al. (2016) observed no differences in

processing between singing and reading aloud during the preparatory phase or otherwise. Furthermore, decrements to performance in Forrin et al. (2019) occurred only for silent items preceding the to-be-produced items, suggesting that participants diverted their attention away from silent items and toward produced items when anticipating the latter. If performance anticipation is greater for singing, this allocation of attention would thereby be expected to *improve* memory for items sung at study at the expense of other conditions. Despite the differential cognitive demands elicited by reading aloud and singing, then, it appears unlikely that an attentional mechanism could obviate the superiority of the latter modality.

Given the boundary conditions that the present study imposes on the singing superiority effect, my 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 effect to be driven by improved memory for singing, my findings cannot rule out the possibility that the effect might be driven by a decrement to performance for aloud items. When colour matching at test was used, numerical trends in my meta-analytic model of the production effect favored lower sensitivity in the aloud condition whilst sensitivity in the sing condition remained comparable across groups (see also, Whitridge, 2022; Experiment 2). Furthermore, I observed no credible difference in sensitivity between conditions in Experiment 2, for which my design did not permit the emergence of within-participant "costs" to performance. Were the singing superiority effect 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.[17] Rather,

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

the pattern of results I 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.

Indeed, earlier research using typical production paradigms has shown that the within-subject production effect arises in part due to performance decrements for silent items (e.g., Bodner et al., 2014) and that adding additional production conditions at study can accentuate this decrement (Ozubko et al., 2020). With respect to my paradigm, it is not immediately clear why participants would prioritize singing over reading aloud, particularly given that Ozubko et al. (2020) only observed decrements to silent items (rather than produced items from other conditions). It could be that preferential attention to sung items could occur due to the inherent peculiarity of the modality relative to reading aloud (but see Quinlan & Taylor, 2019). Alternatively, such a mechanism could arise due to perceived emphasis on singing during instruction or otherwise. In my own investigations, participants were provided with a demonstration of how items should be sung to help ensure task compliance, but the same did not occur for reading aloud; it is possible that such a demonstration could have led some participants to prioritize items sung at study. Analyses of my own data and data provided by others provide some support for this hypothesis: Participants who exhibited a singing superiority effect generally exhibited performance for aloud items that was below average, whilst performance for singing for these participants was near the overall mean. If the singing superiority effect were to be driven by such a mechanism, however, the effect would be attributable to preferential attentional allocation or rehearsal rather than sensorimotor scaling.

The failure of the sensorimotor scaling hypothesis to accommodate my findings implies two possible conclusions, each of which differ in their implications with respect to broader

theoretical frameworks of the production effect. The first possibility is that singing does not encode additional sensorimotor features relative to reading aloud and that the singing superiority effect is driven instead by an alternative mechanism unrelated to encoding distinctiveness (e.g., preferential rehearsal). Were this the case, it would not necessarily imply that the sensorimotor scaling hypothesis is flawed: A great deal of empirical evidence supporting this hypothesis exists (e.g., Conway & Gathercole, 1987; Forrin et al., 2012; Mama & Icht, 2016) and remains unchallenged by the results of my investigation. However, this possibility would imply that the sensorimotor scaling hypothesis can be validated only insofar as removing distinctive processes from the productive act can reduce the size of the production effect. Thus, this conclusion would pose a challenge to the distinctiveness account in that the singing superiority effect does not provide evidence for such a framework. If this conclusion is accepted, whether the sensorimotor scaling hypothesis can be validated bidirectionally remains to be seen, although the failure of repeated efforts to increase the magnitude of the production effect (e.g., Ozubko et al., 2020; Wakeham-Lewis et al., 2022) suggests that this task might prove difficult.[18]

The second possibility is that singing does benefit from additional sensorimotor features, but that these features provide little utility in typical production paradigms. If this conclusion is accepted, it would imply that the singing superiority effect *does* provide solid evidence for distinctiveness accounts, but that the mechanism by which the effect arises is not yet accounted for by such frameworks. In this case, extant formulations of the sensorimotor scaling hypothesis (e.g., Forrin et al., 2012; Mama & Icht, 2016) would require modification: Even if a given modality recruits additional encoding processes at study, participants may not leverage features

[18] Alternatively, the failure of multiple efforts to bidirectionally validate the sensorimotor scaling hypothesis could arise due to a ceiling effect: The production effect is already large, so it may be the case that the benefit cannot be increased further. Were this the case, however, extant formulations of the sensorimotor scaling hypothesis (e.g., Fawcett et al., 2012; Forrin et al., 2012; Mama & Icht, 2016) would nonetheless require modification to accommodate empirical evidence.

related to these processes unless oriented to do so. Accordingly, this conclusion would imply a more nuanced view of distinctiveness in production, wherein the benefit is driven in part by factors beyond sensorimotor features. Although this notion does not pose an inherent challenge to the distinctiveness account, it does argue that the singing superiority effect does not validate the account in the manner that earlier investigations contended (e.g., Quinlan & Taylor, 2013, 2019).

5.4 Implications for The Mnemonic Utility of Singing and Conclusions

Much like investigations within the production literature, broader investigations of singing and memory have produced inconsistent and often conflicting findings (e.g., Kilgour et al., 2000; Rainey & Larsen, 2002; Salcedo, 2010; Wallace, 1994). With respect to the present investigation, my findings provide mixed evidence for the utility of singing as a mnemonic aid. First, the present thesis consistently observed a large production effect for singing, which meta-analysis revealed to be robust across studies. Consistent with earlier research, then, my findings suggest that singing almost certainly improves item memory relative to silent reading (e.g., Hassall et al., 2016; Quinlan & Taylor, 2013, 2019; Whitridge, 2022). Further, Experiment 2 is the first investigation to show that the production effect for singing emerges in between-subject designs, suggesting that the benefit is absolute and not relative (i.e., being driven by a cost to silent items; e.g., Bodner et al., 2014). Accordingly, I can reasonably conclude that singing is a powerful mnemonic and appears to improve item memory at least as much as reading aloud.

However, my findings do not provide strong support for the notion that singing possesses superior mnemonic utility relative to reading aloud. Based on Experiment 1, it appears that the singing superiority effect is replicable. However, it is important to consider not only whether singing can boost memory further, but whether this increase in performance can be leveraged in

a meaningful way. To this point, both Experiment 1 and my meta-analytic models strongly suggested that the singing superiority effect appeared to be driven by color matching at test. While it appears that singing can produce superior memory, then, this additional benefit emerges only when a specific, atypical set of conditions are met. Thus, the utility of singing as a superior mnemonic and the generalizability of this finding to other designs is dubious. Furthermore, given that Experiment 2 failed to detect a singing superiority effect, the benefit appears to occur only relative to aloud items and may emerge due to preferential rehearsal or an accentuated decrement to silent items (see, e.g., Forrin et al., 2019; Ozubko et al., 2020). Accordingly, the applicability of this strategy in other contexts might be limited: If singing is not an inherently superior modality, it would appear to possess little utility over reading aloud for improving learning in comprehension or second language acquisition paradigms, where the modality has received attention (e.g., Baills et al., 2021; Ludke, 2014). While it is possible that the study design or materials used in investigations outside the production literature might be more conducive to the emergence of singing superiority effects, the evidence available at present leads me to conclude that the mnemonic utility of singing is probably not meaningfully different from reading aloud.

In sum, the present thesis poses a challenge to the singing superiority effect as described in earlier literature (e.g., Quinlan & Taylor, 2013, 2019). Across several analyses, I observed a production effect for singing that was generally similar in magnitude to that for reading aloud (see also, Hassall et al., 2016). When the singing superiority effect 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 likely arises due to idiosyncratic methodological factors. Even if these factors can be leveraged via an atypical distinctiveness heuristic or some alternative, context-based 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 singing superiority effect does not appear to arise solely on the basis of appending additional distinctive features to the production trace, the present thesis argues that the effect should not be construed as strong evidence for the sensorimotor scaling hypothesis and, by extension, the distinctiveness account.

References

Anton, R. J. (1990). Combining singing and psychology. *Hispania*, *73*(4), 1166-1170. https://doi.org/10.2307/344326

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

Bailey, L. M., Bodner, G. E., Matheson, H. E., Stewart, B. M., Roddick, K., O'Neil, K., Simmons, M., Lambert, A. M., Krigolson, O. E., Newman, A. J., & Fawcett, J. M. (2021). Neural correlates of the production effect: An fMRI study. *Brain and Cognition, 152,* Article 105757. https://doi.org/10.1016/j.bandc.2021.105757

Baills, F., Zhang, Y., Cheng, Y., Bu, Y., & Prieto, P. (2021). Listening to songs and singing benefitted initial stages of second language pronunciation but not recall of word meaning. *Language Learning, 71*(2), 369–413. https://doi.org/10.1111/lang.12442

Banbury, S. P., Tremblay, S., Macken, W. J., & Jones, D. M. (2001). Auditory distraction and short-term memory: Phenomena and practical implications. *Human Factors, 43*(1), 12–29. https://doi.org/10.1518/001872001775992462

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

Beatty, J. (1982). Task-evoked pupillary responses, processing load, and the structure of processing resources. *Psychological Bulletin, 91*(2), 276–292. https://doi.org/10.1037/0033-2909.91.2.276

Beaty, R. E. (2015). The neuroscience of musical improvisation. *Neuroscience and Biobehavioral Reviews, 51,* 108–117. https://doi.org/10.1016/j.neubiorev.2015.01.004

Benjamin, A. S., Tullis, J. G., & Lee, J. H. (2013). Criterion noise in ratings-based recognition: Evidence from the effects of response scale length on recognition accuracy. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 39*(5), 1601–1608. https://doi.org/10.1037/a0031849

Berger, A., & Kiefer, M. (2021). Comparison of different response time outlier exclusion methods: A simulation study. *Frontiers in Psychology, 12,* Article 675558. https://doi.org/10.3389/fpsyg.2021.675558

Bodner, G. E., Huff, M. J., & Taikh, A. (2020). Pure-list production improves item recognition and sometimes also improves source memory. *Memory & Cognition, 48*(7), 1281–1294. https://doi.org/10.3758/s13421-020-01044-2

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. (2012). Reassessing the basis of the production effect in memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 38*(6), 1711–1719. https://doi.org/10.1037/a0028466

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

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

Borenstein, M., & Higgins, J. P. (2013). Meta-analysis and subgroups. *Prevention Science*, *14*, 134-143. https://doi.org/10.1007/s11121-013-0377-7

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

Burke, J., & Ornstein, R. (2018). Communication and faith in the Middle Ages. In *Communication in History* (pp. 65-71). Routledge.

Bürkner, P. C. (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

Bürkner, P. C. (2021). Bayesian Item Response Modeling in R with brms and Stan. *Journal of Statistical Software*, *100*, 1-54. https://doi.org/10.18637/jss.v100.i05

Burrows, D. (1989). Singing and saying. *The Journal of Musicology*, *7*(3), 390-402. https://doi.org/10.2307/763607

Castel, A. D., Rhodes, M. G., & Friedman, M. C. (2013). Predicting memory benefits in the production effect: The use and misuse of self-generated distinctive cues when making judgments of learning. *Memory & Cognition, 41*(1), 28–35. https://doi.org/10.3758/s13421-012-0249-6

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

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

Crowder, R. G. (1970). The role of one's own voice in immediate memory. *Cognitive Psychology*, *1*(2), 157-178. https://doi.org/10.1016/0010-0285(70)90011-3

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

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

Dixon, P. (2008). Models of accuracy in repeated-measures designs. *Journal of Memory and Language, 59*(4), 447–456. https://doi.org/10.1016/j.jml.2007.11.004

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

Donchin, E. (1981). Surprise!… Surprise? *Psychophysiology*, *18*(5), 493-513. https://doi.org/10.1111/j.1469-8986.1981.tb01815.x

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

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

Einstein, G. O., McDaniel, M. A., & Lackey, S. (1989). Bizarre imagery, interference, and distinctiveness. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 15*(1), 137–146. https://doi.org/10.1037/0278-7393.15.1.137

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

Fawcett, J. M., Bodner, G. E., Paulewicz, B., Rose, J., & Wakeham-Lewis, R. (2022b). Production can enhance semantic encoding: Evidence from forced-choice recognition with homophone versus synonym lures. *Psychonomic Bulletin & Review*, *29*(6), 2256-2263. https://doi.org/10.3758/s13423-022-02140-x

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

Feltgen, Q., & Daunizeau, J. (2021). An overcomplete approach to fitting drift-diffusion decision models to trial-by-trial data. *Frontiers in Artificial Intelligence*, *4*, 531316. https://doi.org/10.3389/frai.2021.531316

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

Forrin, N. D., Ralph, B. C. W., Dhaliwal, N. K., Smilek, D., & MacLeod, C. M. (2019). Wait for it… Performance anticipation reduces recognition memory. *Journal of Memory and Language, 109,* Article 104050. https://doi.org/10.1016/j.jml.2019.104050

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

Geiser, E., Zaehle, T., Jancke, L., & Meyer, M. (2008). The neural correlate of speech rhythm as evidenced by metrical speech processing. *Journal of cognitive neuroscience*, *20*(3), 541-552. https://doi.org/10.1162/jocn.2008.20029

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

Gfeller, K. E. (1983). Musical mnemonics as an aid to retention with normal and learning disabled students. *Journal of Music Therapy*, *20*(4), 179-189. https://doi.org/10.1093/jmt/20.4.179

Gionet, S., Guitard, D., & Saint-Aubin, J. (2022). The production effect interacts with serial positions: Further evidence from a between-subjects manipulation. *Experimental Psychology, 69*(1), 12–22. https://doi.org/10.1027/1618-3169/a000540

Gionet, S., Guitard, D., & Saint-Aubin, J. (in press). The production effect interacts with serial positions in recall tasks, but not in item recognition. *Experimental Psychology*.

Good, A. J., Russo, F. A., & Sullivan, J. (2015). The efficacy of singing in foreign-language learning. *Psychology of Music, 43*(5), 627–640. https://doi.org/10.1177/0305735614528833

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

Greene, R. L., & Crowder, R. G. (1984). Modality and suffix effects in the absence of auditory stimulation. *Journal of Verbal Learning & Verbal Behavior, 23*(3), 371–382. https://doi.org/10.1016/S0022-5371(84)90259-7

Gretz, M. R., & Huff, M. J. (2020). Multiple species of distinctiveness in memory? Comparing encoding versus statistical distinctiveness on recognition. *Memory, 28*(8), 984–997. https://doi.org/10.1080/09658211.2020.1803916

Harrer, M., Cuijpers, P., Furukawa, T.A., & Ebert, D.D. (2021). *Doing meta-analysis with R: A hands-on guide*. 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

Hintzman, D. L. (1984). MINERVA 2: A simulation model of human memory. *Behavior Research Methods, Instruments & Computers, 16*(2), 96–101. https://doi.org/10.3758/BF03202365

Hofmann, S. G., Moscovitch, D. A., & Kim, H.-J. (2006). Autonomic correlates of social anxiety and embarrassment in shy and non-shy individuals. *International Journal of Psychophysiology, 61*(2), 134–142. https://doi.org/10.1016/j.ijpsycho.2005.09.003

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

Hourihan, K. L., & Churchill, L. A. (2020). Production of picture names improves picture recognition. *Canadian Journal of Experimental Psychology, 74*(1), 35–43. https://doi.org/10.1037/cep0000185

Huff, M. J., Bodner, G. E., & Fawcett, J. M. (2015). Effects of distinctive encoding on correct and false memory: A meta-analytic review of costs and benefits and their origins in the DRM paradigm. *Psychonomic Bulletin & Review, 22*(2), 349–365. https://doi.org/10.3758/s13423-014-0648-8

Hunt, R. R. (2006). The concept of distinctiveness in memory research. In R. R. Hunt & J. B. Worthen (Eds.), *Distinctiveness and memory* (pp. 3–25). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195169669.003.0001

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

Icht, M., & Mama, Y. (2015). The production effect in memory: A prominent mnemonic in children. *Journal of Child Language, 42*(5), 1102–1124. https://doi.org/10.1017/S0305000914000713

Icht, M., Mama, Y., & Algom, D. (2014). The production effect in memory: Multiple species of distinctiveness. *Frontiers in Psychology, 5,* Article 886. https://doi.org/10.3389/fpsyg.2014.00886

Icht, M., Taitelbaum-Swead, R., & Mama, Y. (2022). Production improves visual and auditory text memory in younger and older adults. *Gerontology, 68*(5), 578–586. https://doi.org/10.1159/000518894

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

Isreal, J. B., Chesney, G. L., Wickens, C. D., & Donchin, E. (1980). P300 and tracking difficulty: Evidence for multiple resources in dual-task performance. *Psychophysiology, 17*(3), 259–273. https://doi.org/10.1111/j.1469-8986.1980.tb00146.x

Jacoby, L. L. (1983). Remembering the data: Analyzing interactive processes in reading. *Journal of Verbal Learning & Verbal Behavior, 22*(5), 485–508. https://doi.org/10.1016/S0022-5371(83)90301-8

Jacoby, L. L., Woloshyn, V., & Kelley, C. (1989). Becoming famous without being recognized: Unconscious influences of memory produced by dividing attention. *Journal of Experimental Psychology: General, 118*(2), 115–125. https://doi.org/10.1037/0096-3445.118.2.115

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

James, W. (1890). *The principles of psychology*. New York: Henry Holt and Company. https://doi.org/10.1037/11059-000

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

Jamieson, R. K., & Spear, J. (2014). The offline production effect. *Canadian Journal of Experimental Psychology, 68*(1), 20–28. https://doi.org/10.1037/cep0000009

Jeffries, K. J., Fritz, J. B., & Braun, A. R. (2003). Words in melody: An $H_2{}^{15}O$ PET study of brain activation during singing and speaking. *NeuroReport: For Rapid Communication of Neuroscience Research, 14*(5), 749–754. https://doi.org/10.1097/00001756-200304150-00018

Johnson, M. K., Hashtroudi, S., & Lindsay, D. S. (1993). Source monitoring. *Psychological Bulletin, 114*(1), 3–28. https://doi.org/10.1037/0033-2909.114.1.3

Jones, A. C., & Pyc, M. A. (2014). The production effect: Costs and benefits in free recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 40*(1), 300–305. https://doi.org/10.1037/a0033337

Jones, R. M., Fox, R. A., & Jacewicz, E. (2012). The effects of concurrent cognitive load on phonological processing in adults who stutter. *Journal of Speech, Language, and Hearing Research, 55*(6), 1862–1875. https://doi.org/10.1044/1092-4388(2012/12-0014)

Jonker, T. R., Levene, M., & MacLeod, C. M. (2014). Testing the item-order account of design effects using the production effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 40*(2), 441–448. https://doi.org/10.1037/a0034977

Kappel, S., Harford, M., Burns, V. D., & Anderson, N. S. (1973). Effects of vocalization on short-term memory for words. *Journal of Experimental Psychology, 101*(2), 314–317. https://doi.org/10.1037/h0035247.

Karis, D., Fabiani, M., & Donchin, E. (1984). "P300" and memory: Individual differences in the von Restorff effect. *Cognitive Psychology, 16*(2), 177–216. https://doi.org/10.1016/0010-0285(84)90007-0

Kellogg, R. T., Newcombe, C., Kammer, D., & Schmitt, K. (1996). Attention in direct and indirect memory tasks with short- and long-term probes. *The American Journal of Psychology, 109*(2), 205–217. https://doi.org/10.2307/1423273

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

Kilgour, A. R., Jakobson, L. S., & Cuddy, L. L. (2000). Music training and rate of presentation as mediators of text and song recall. *Memory & Cognition, 28*(5), 700–710. https://doi.org/10.3758/BF03198404

Kok, A. (2001). On the utility of P3 amplitude as a measure of processing capacity. *Psychophysiology*, *38*(3), 557-577. https://doi.org/10.1017/S0048577201990559

Kolinsky, R., Lidji, P., Peretz, I., Besson, M., & Morais, J. (2009). Processing interactions between phonology and melody: Vowels sing but consonants speak. *Cognition, 112*(1), 1–20. https://doi.org/10.1016/j.cognition.2009.02.014

Kramer, A. F., Wickens, C. D., & Donchin, E. (1983). An analysis of the processing requirements of a complex perceptual-motor task. *Human Factors, 25*(6), 597–621. https://doi.org/10.1177/001872088302500601

Kruschke, J. K. (2010). Bayesian data analysis. *Wiley Interdisciplinary Reviews: Cognitive Science*, *1*(5), 658-676. https://doi.org/10.1002/wcs.72

Kruschke, J. (2014). *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan*. Academic Press.

Lambert, A. M., Bodner, G. E., & Taikh, A. (2016). The production effect in long-list recall: In no particular order? *Canadian Journal of Experimental Psychology, 70*(2), 165–176. https://doi.org/10.1037/cep0000086

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

Lentz, T. M. (1985). From recitation to reading: Memory, writing, and composition in Greek philosophical prose. *Southern Speech Communication Journal*, *51*(1), 49-70. https://doi.org/10.1080/10417948509372646

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

Lloyd, M. E., & Miller, J. K. (2011). Are two heuristics better than one? The fluency and distinctiveness heuristics in recognition memory. *Memory & Cognition, 39*(7), 1264–1274. https://doi.org/10.3758/s13421-011-0093-0

Ludke, K. M., Ferreira, F., & Overy, K. (2014). Singing can facilitate foreign language learning. *Memory & Cognition, 42*(1), 41–52. https://doi.org/10.3758/s13421-013-0342-5

MacDonald, P. A., & MacLeod, C. M. (1998). The influence of attention at encoding on direct and indirect remembering. *Acta Psychologica*, *98*(2-3), 291-310. https://doi.org/10.1016/S0001-6918(97)00047-4

MacLeod, C. M., & Bodner, G. E. (2017). The production effect in memory. *Current Directions in Psychological Science, 26*(4), 390–395. https://doi.org/10.1177/0963721417691356

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

MacLeod, C. M., Ozubko, J. D., Hourihan, K. L., & Major, J. C. (2022). The production effect is consistent over material variations: Support for the distinctiveness account. *Memory*, *30*(8), 1000-1007. https://doi.org/10.1080/09658211.2022.2069270

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

Mama, Y., Fostick, L., & Icht, M. (2018). The impact of different background noises on the production effect. *Acta Psychologica, 185,* 235–242. https://doi.org/10.1016/j.actpsy.2018.03.002

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

Mama, Y., & Icht, M. (2018). Production on hold: Delaying vocal production enhances the production effect in free recall. *Memory, 26*(5), 589–602. https://doi.org/10.1080/09658211.2017.1384496

Mama, Y., & Icht, M. (2019). Production effect in adults with ADHD with and without methylphenidate (MPH): Vocalization improves verbal learning. *Journal of the International Neuropsychological Society, 25*(2), 230–235. https://doi.org/10.1017/S1355617718001017

Marshall, P. H., & Werder, P. R. (1972). The effects of the elimination of rehearsal on primacy and recency. *Journal of Verbal Learning & Verbal Behavior, 11*(5), 649–653. https://doi.org/10.1016/S0022-5371(72)80049-5

Masson, M. E. J. (2011). A tutorial on a practical Bayesian alternative to null-hypothesis significance testing. *Behavior Research Methods, 43*(3), 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

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

McGettigan, C., Eisner, F., Agnew, Z. K., Manly, T., Wisbey, D., & Scott, S. K. (2013). T'ain't what you say, it's the way that you say it—left insula and inferior frontal cortex work in interaction with superior temporal regions to control the performance of vocal impersonations. *Journal of Cognitive Neuroscience*, *25*(11), 1875-1886. https://doi.org/10.1162/jocn_a_00427

Morehead, K., Rhodes, M. G., & DeLozier, S. (2016). Instructor and student knowledge of study strategies. *Memory*, *24*(2), 257-271. https://doi.org/10.1080/09658211.2014.1001992

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

Mulligan, N. W. (2011). Generation disrupts memory for intrinsic context but not extrinsic context. *The Quarterly Journal of Experimental Psychology, 64*(8), 1543–1562. https://doi.org/10.1080/17470218.2011.562980

Mulligan, N. W., & Hartman, M. (1996). Divided attention and indirect memory tests. *Memory & Cognition, 24*(4), 453–465. https://doi.org/10.3758/BF03200934

Nairne, J. S. (1990). A feature model of immediate memory. *Memory & Cognition, 18*(3), 251–269. https://doi.org/10.3758/BF03213879

Namias, J. M., Huff, M. J., Smith, A., & Maxwell, N. P. (2022). Drawing individual images benefits recognition accuracy in the Deese–Roediger–McDermott paradigm. *The Quarterly Journal of Experimental Psychology, 75*(8), 1571–1582. https://doi.org/10.1177/17470218211056498

Otten, L. J., & Donchin, E. (2000). Relationship between P300 amplitude and subsequent recall for distinctive events: Dependence on type of distinctiveness attribute. *Psychophysiology, 37*(5), 644–661. https://doi.org/10.1017/S004857720098171X

Özdemir, E., Norton, A., & Schlaug, G. (2006). Shared and distinct neural correlates of singing and speaking. *Neuroimage*, *33*(2), 628-635. https://doi.org/10.1016/j.neuroimage.2006.07.013

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., Hourihan, K. L., & MacLeod, C. M. (2012a). Production benefits learning: The production effect endures and improves memory for text. *Memory, 20*(7), 717–727. https://doi.org/10.1080/09658211.2012.699070

Ozubko, J. D., Gopie, N., & MacLeod, C. M. (2012b). 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

Ozubko, J. D., Major, J., & MacLeod, C. M. (2014). Remembered study mode: Support for the distinctiveness account of the production effect. *Memory, 22*(5), 509–524. https://doi.org/10.1080/09658211.2013.800554

Paivio, A. (1991). *Images in mind: The evolution of a theory.* Harvester Wheatsheaf.

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

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

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

Polich, J., & Kok, A. (1995). Cognitive and biological determinants of P300: An integrative review. *Biological Psychology, 41*(2), 103–146. https://doi.org/10.1016/0301-0511(95)05130-9

Pressing, J. (1988). Improvisation: Methods and models. In J. A. Sloboda (Ed.), *Generative processes in music: The psychology of performance, improvisation, and composition* (pp. 129–178). Clarendon Press/Oxford University Press.

Prickett, C. A., & Moore, R. S. (1991). The use of music to aid memory of Alzheimer's patients. *Journal of Music Therapy, 28*(2), 101–110. https://doi.org/10.1093/jmt/28.2.101

Pritchard, V. E., Heron-Delaney, M., Malone, S. A., & MacLeod, C. M. (2020). The production effect improves memory in 7- to 10-year-old children. *Child Development, 91*(3), 901–913. https://doi.org/10.1111/cdev.13247

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 effects 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

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review, 85*(2), 59–108. https://doi.org/10.1037/0033-295X.85.2.59

Ratcliff, R., Gomez, P., & McKoon, G. (2004). A Diffusion Model Account of the Lexical Decision Task. *Psychological Review, 111*(1), 159–182. https://doi.org/10.1037/0033-295X.111.1.159

Ratcliff, R., & Rouder, J. N. (1998). Modeling response times for two-choice decisions. *Psychological Science*, *9*(5), 347-356. https://doi.org/10.1111/1467-9280.00067

Ratcliff, R., Voskuilen, C., & Teodorescu, A. (2018). Modeling 2-alternative forced-choice tasks: Accounting for both magnitude and difference effects. *Cognitive Psychology, 103,* 1–22. https://doi.org/10.1016/j.cogpsych.2018.02.002

Richards, J. (1969). Songs in language learning. *Tesol Quarterly*, *3*(2), 161-174. https://doi.org/10.2307/3586103

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

Riefer, D. M., Chien, Y., & Reimer, J. F. (2007). Positive and negative generation effects in source monitoring. *The Quarterly Journal of Experimental Psychology, 60*(10), 1389–1405. https://doi.org/10.1080/17470210601025646

Robinson, J. A., & Taylor, L. R. (2014). Autobiographical memory and self-narratives: A tale of two stories. In *Autobiographical memory* (pp. 125-143). Psychology Press.

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

Routh, D. A. (1970). 'Trace strength,' modality, and the serial position curve in immediate memory. *Psychonomic Science*, *18*(6), 355-357. https://doi.org/10.3758/BF03332397

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

Salcedo, C. S. (2010). The effects of songs in the foreign language classroom on text recall, delayed text recall and involuntary mental rehearsal. *Journal of College Teaching & Learning*, *7*(6).

Schacter, D. L., Israel, L., & Racine, C. (1999). Suppressing false recognition in younger and older adults: The distinctiveness heuristic. *Journal of Memory and Language, 40*(1), 1–24. https://doi.org/10.1006/jmla.1998.2611

Shafritz, K. M., Marchione, K. E., Gore, J. C., Shaywitz, S. E., & Shaywitz, B. A. (2004). The effects of methylphenidate on neural systems of attention in attention deficit hyperactivity disorder. *The American Journal of Psychiatry, 161*(11), 1990–1997. https://doi.org/10.1176/appi.ajp.161.11.1990

Singmann, H. (2017, November 26). Diffusion/Wiener model analysis with brms – Part I: Introduction and estimation. http://singmann.org/wiener-model-analysis-with-brms-part-i/

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

Taikh, A., & Bodner, G. E. (2016). Evaluating the basis of the between-group production effect in recognition. *Canadian Journal of Experimental Psychology, 70*(2), 186–194. https://doi.org/10.1037/cep0000083

Taitelbaum-Swead, R., Icht, M., & Mama, Y. (2017). The effect of learning modality and auditory feedback on word memory: Cochlear-implanted versus normal-hearing adults. *Journal of the American Academy of Audiology*, *28*(03), 222-231. https://doi.org/10.3766/jaaa.16032

Taitelbaum-Swead, R., 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

Varao Sousa, T. L., Carriere, J. S., & Smilek, D. (2013). The way we encounter reading material influences how frequently we mind wander. *Frontiers in Psychology*, *4*, 892. https://doi.org/10.3389/fpsyg.2013.00892

Vehtari, A., Gabry, J., Magnusson, M., Yao, Y., Bürkner, P., Paananen, T., & Gelman, A. (2024). loo: Efficient leave-one-out cross-validation and WAIC for Bayesian models. R package version 2.7.0. https://mc-stan.org/loo/

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, *27*, 1413-1432. https://doi.org/10.1007/s11222-016-9696-4

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

Voss, A., Nagler, M., & Lerche, V. (2013). Diffusion models in experimental psychology: A practical introduction. *Experimental Psychology, 60*(6), 385–402. https://doi.org/10.1027/1618-3169/a000218

Voss, A., Voss, J., & Lerche, V. (2015). Assessing cognitive processes with diffusion model analyses: A tutorial based on fast-dm-30. *Frontiers in Psychology*, *6*, 124917. https://doi.org/10.3389/fpsyg.2015.00336

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/

Wagenmakers, E. J. (2009). Methodological and empirical developments for the Ratcliff diffusion model of response times and accuracy. *European Journal of Cognitive Psychology, 21*(5), 641–671. https://doi.org/10.1080/09541440802205067

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

Wiecki, T. V., Sofer, I., & Frank, M. J. (2013). HDDM: Hierarchical Bayesian estimation of the drift-diffusion model in Python. *Frontiers in Neuroinformatics*, *7*, 55610. https://doi.org/10.3389/fninf.2013.00014

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

Whitehead, J. C., & Armony, J. L. (2018). Singing in the brain: Neural representation of music and voice as revealed by fMRI. *Human Brain Mapping, 39*(12), 4913–4924. https://doi.org/10.1002/hbm.24333

Whitridge, J. W. (2022) *Does the song remain the same? Singing does not necessarily improve memory more than reading aloud* [Undergraduate honours thesis]. Memorial University of Newfoundland and Labrador.

Whitridge, J. W., Clark, C. A., Hourihan, K. L., & Fawcett, J. M. (2024). *Generation (not production) improves the fidelity of visual representations in picture naming*. [Manuscript submitted for publication].

Willoughby, H. V., Tiller, J., Hourihan, K. L., & Fawcett, J. M. (2019). *The pupillometric production effect: Measuring attentional engagement during a production task* [Paper presentation]. CSBBCS 2019 Meeting, Waterloo, Canada.

Wilson, M., & Emmorey, K. (1997). Working memory for sign language: A window into the architecture of the working memory system. *Journal of Deaf Studies and Deaf Education, 2*(3), 121–130. https://doi.org/10.1093/oxfordjournals.deafed.a014318

Wood, S.N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society*, *73*, 3-36. https://doi.org/10.1111/j.1467-9868.2010.00749.x

Wood, S. N. (2017). *Generalized additive models: An introduction with R*. Chapman and Hall/CRC. https://doi.org/10.1201/9781315370279

Woods, S. P., Lovejoy, D. W., & Ball, J. D. (2002). Neuropsychological characteristics of adults with ADHD: A comprehensive review of initial studies. *The Clinical Neuropsychologist, 16*(1), 12–34. https://doi.org/10.1076/clin.16.1.12.8336

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

Zatorre, R. J., & Baum, S. R. (2012). Musical melody and speech intonation: Singing a different tune. *PLoS Biol*, *10*(7), e1001372. https://doi.org/10.1371/journal.pbio.1001372

Zhang, B. (2024). *Comparing memory levels between reading aloud and singing*. Unpublished manuscript.

Zhang, B., Meng, Z., Li, Q., Chen, A., & Bodner, G. E. (2023). EEG-based univariate and multivariate analyses reveal that multiple processes contribute to the production effect in recognition. *Cortex, 165,* 57–69. https://doi.org/10.1016/j.cortex.2023.04.006

Zhou, Y., & MacLeod, C. M. (2021). Production between and within: Distinctiveness and the relative magnitude of the production effect. *Memory, 29*(2), 168–179. https://doi.org/10.1080/09658211.2020.1868526

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