Optimizing Song Retention Through Spacing

JOEL KATZ

A DISSERTATION SUBMITTED TO

THE FACULTY OF GRADUATE STUDIES

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

GRADUATE PROGRAM IN MUSIC

YORK UNIVERSITY

TORONTO, ONTARIO

May 2019

Abstract

The distributed practice or spacing effect refers to the improvement in memory retention for materials learned in a series of distributed sessions over learning massed in a single session. It has been studied in the domains of verbal learning, motor skill learning and complex co-ordination. The effect of distributed learning on the retention of words and music in song, however, has yet to be determined. This dissertation examines the effect of different spacing intervals on song retention among a population of university undergraduates. A group of second year undergraduate music students (n = 70) supplemented by a group of university students from the general population (n = 17) learned an unaccompanied two-verse song based on traditional materials to a criterion of 95% correct memory for sung words. Subsequent training sessions were spaced at intervals of ten minutes, two days or one week, and tested at a retention interval of three weeks. Performances were evaluated for word errors, additions and omissions; pitch accuracy and omissions; and hesitation as measured by mean note length, length of breaths and length of hesitations. After the third session all participants were tested for musical perception skills using tests drawn from the shorter PROMS battery. Results were analyzed with Bayesian ANOVA and additional post hoc tests. The data revealed very strong evidence for a spacing effect for song between the massed (ten-minute gap) and spaced conditions at a retention interval of three weeks, and evidence of no difference between the two spaced conditions, with large effect sizes for syllable memory (d = 0.873; d = 0.914). These findings suggest that the ongoing cues offered from surface features in the song are strong enough to enable verbatim recall across spaced conditions, as long as the spacing interval reaches a critical threshold.

Keywords: Distributed practice effect song, distributed practice effect music, song spacing effect, music spacing effect, song memory, song learning.

Acknowledgements

I would like to thank my dissertation supervisors, Dr. Dorothy DeVal and Dr. Melody Wiseheart for their support and encouragement, without which this dissertation never would have been completed. Dr. Wiseheart was an unfailing and on-going source of advice, expertise and inspiration throughout the data collection and writing up of the results. In addition, Dr. Wiseheart wrote the software used for analysis and ran the data analyses for this dissertation. Dr. DeVal's help was crucial in gaining access to the pool of music students on which this research is primarily based. Irene Backhouse at the LaMarsh Centre for Child and Youth Research very kindly donated laboratory space to run the experiment and was an ongoing source of practical advice and encouragement. The Cognitive Flexibility Lab at York was a wonderful home for the research and offered invaluable comments throughout the lengthy dissertation process. The three undergraduates who helped with data transcription, Momo Ando, Isabella Bozzo and Alex Tran, were a huge help. I would also like to thank my other committee members, Dr. Sherry Johnson and Dr. Gary Turner, and the SSHRC for a three-year Fellowship. I am very grateful to York University, who in addition to providing ongoing funding for the PHD program, awarded a Provost's Scholarship to me in my final year. This facilitated completion of the dissertation in my fifth year of PhD studies. Finally, my long-suffering companion, Tim Montgomery, has put up with my doubts and struggles throughout this entire five-year period. Without his love and support this work would never have been completed.

Table of Contents

List of Tables

List of Figures

1: Introduction

1.1 Objectives

This research is a study of spacing effects in singers' memory. The *spacing effect* is an observable boost to memory performance found when learning is distributed over a number of different sessions compared to learning in a single session. It has been studied since the time of the pioneering psychologist, Hermann Ebbinghaus (1885/1913), often using word lists and simple cognitive tasks (Cepeda, Pashler, Vul, Vixted & Rohrer, 2006).[1] Complex tasks, however, have been evaluated much less frequently. Singing, with its unique combination of words and music, has never been evaluated for the optimal distribution of learning events over time. This study is an initial effort to address that gap in the literature.

The study has practical implications for singers, presenting evidence for learning schedules that can best support performance memory. It may be of use to teachers of musical mnemonics, a rehabilitative program for verbal memory deficits (Murakami, 2019), and is relevant to foreign language instruction (Ludke, 2018; Good, Russo & Sullivan, 2014). More generally, this study may be of interest to any educator designing a program that utilizes song as an aid to memory. The results are consistent with a theory of multiple constraints in the operation of song memory (Rubin, 1997) and the multiple systems theory of episodic memory (Rubin, 2006).

1.2 Background of the study

Song is found everywhere in human culture. It has been used to transmit ancient texts verbatim across millennia (Rubin, 1997). Long before the advent of written language, specialist

---

[1] Ebbinghaus was the first to introduce the *forgetting curve*, a plot of the number of trials necessary to return to an error-free performance after an interval.

singers learned to deliver thousands of notes and words in a time-limited fluent manner, a process that continues today (Live Nation Entertainment, 2018; English, 2007; Sgroi, 2005)[2]. In contemporary life, song has been shown to have significance for establishing self-concept and for emotional regulation (DeNora, 2000). Song "brings people together in the expression of common sentiments and promotes the solidarity of groups" (Byron, 1995, p. 14). Singing has been used as an aid to language instruction (Ludke, 2018; Good, Russo & Sullivan, 2014), in the rehabilitation of language disorders (Aldridge, 2005) and as a mnemonic in medical education (Cirigliano, 2013). Music may act as a cue to reminiscence (Sloboda, 1999), and song in particular is a "valuable cue to evoke autobiographical memory" (Cady, Harris & Knappenberger, 2008, p. 157). Given the ubiquity and importance of song, it would be helpful to know what approaches to practice might enhance the long-term retention of vocal materials.

The spacing of learning sessions is a well-established tool for improving verbal learning (Carpenter, Cepeda, Rohrer, Kang & Pashler, 2012; Wiseheart, Küpper-Tetzel, Weston, Kim, Kapler & Foot, 2019). Where the time between the final learning session and the test is more than a certain critical interval (varying with the choice of materials), spaced practice (distributed learning) will be more effective than massed practice for verbal retention. This boost to memory is called the *spacing effect*. There are only seven studies in the literature evaluating distributed learning for music, and with one exception, most have been published fairly recently (see Cash, 2009; Rubin-Rabson, 1940; Simmons & Duke, 2006; Simmons, 2012; Stambaugh, 2011;

---

[2] For example, the average number of words in a song from the album "Born to be Wild" by Bruce Springsteen is 281 (English, http://i.grahamenglish.net/1163/average-words-per-song-and-the-8020-rule/). The number of songs in a Springsteen set list from 2018 averages 15 (Live Nation Entertainment https://www.setlist.fm/stats/average-setlist/bruce-springsteen-2bd6dcce.html?year=2018). Clearly, there are thousands of words to be learned and memorized in any solo performance.

.

Stambaugh & Demorest, 2010; and Wiseheart, D'Souza & Chae, 2017). For singing, with the added cognitive demands of accurate, rhythm-limited textual recall, there are no studies of distributed learning.

This study is the first to look at the interaction of song with spacing as an aid to song learning. It is a quantitative evaluation of errors in syllables, pitch and rhythm for a two-verse song based on traditional materials, taught under a variety of spaced learning conditions, and evaluated after a three-week interval. The independent variables were the three learning conditions, with a ten-minute, a two-day or a one-week interval between two rehearsals. Dependent variables included syllable, pitch and rhythm errors at various tests during the sessions. Demographic factors and a perceptual musical skills battery (the PROMS test) were also tracked to establish equivalence across conditions.

There are two main theories from the spacing literature and one general theory of episodic memory which have served as a basis for the hypotheses in this study. Explanations of the spacing effect have focused on *contextual variability* and *study-phase retrieval* as the main drivers of the effect (Glenberg, 1979; Delaney, Verkoejen & Spirgel, 2010; Küpper-Tetzel, Kapler & Wiseheart, 2014). According to contextual variability theory, every learning session has its associated context, consisting largely of unconscious elements that describe the learning environment (Glenberg, 1979). "Contextual variability" refers to the differing kind and number of contextual cues which may be encoded in different learning events. Separating learning sessions may add to those cues, enhancing the likelihood of overlap between stored contextual cues and contextual cues available at retrieval. "Study-phase retrieval" refers to the retrieval of learned information at the second testing session. Where the attempt at retrieval is more difficult (after a space between learning sessions, for instance) retrieving the memory trace during learning will strengthen the memory. This results in improved performance on a final memory test after a retention interval. If the space between learning sessions is too short, retrieval is too easy and little reconstruction (and thus, strengthening) of the initial memory trace occurs. Performance on the final test will suffer as a result. Similarly, when intervals between study sessions are too long, the initial memory trace may be forgotten, and the second learning event coded as a separate event. Performance on the final test will then be less successful.

Glenberg (1979) demonstrated that for relatively simple verbal materials there is an optimal gap between learning sessions, beyond which the spacing effect diminishes. The optimal gap varies depending on the nature of the materials and the length of the retention interval. For more complex materials, however, the learning profile may be quite different (Wulf & Shea, 2002). It is not unreasonable, however, to suppose that for any given retention interval there may be a gap between song learning sessions that is enough to enhance the performance of a song from memory; and not so much that the song is forgotten.

The multiple systems theory of episodic memory predicts that where more basic systems are involved in coding, memory retrieval will be improved (Rubin, 2006). A song based on traditional materials such as folksongs or ballads may introduce cues related to multiple episodic systems. This type of song contains a wealth of associative elements related to ongoing constraints from the prosodic structure, imagery, rhyme, and metre of the text; the metrical and pitch characteristics of the musical setting; the narrative and affective nature of sung material; and the proprioceptive aspects of singing (Wallace & Rubin, 1991). These strongly associative cues generated within the song materials may be enough to overtake any additional boost to learning offered by both increased contextual variability and increased study-phase retrieval at increasing spaced intervals. For this reason, there should be minimal difference in error scores for the syllables of a traditional song between two spaced learning episodes, when compared at the same retention interval.

Forgetting, as measured at the start of the second learning session, should show the effect of both contextual variability and the song constraints recovered from the cue. The word and note errors in the massed condition should be low, because the retrieval context at the beginning of the second session will be a close match to the learning context. That, combined with the structural information retrieved from the cue (which acts to constrain note and word choice) and the ongoing associative cues generated within the songs themselves, should produce excellent retrieval for the massed condition. At the spaced intervals, the mismatch in learning context between the first and second session and forgetting of the initial memory trace should produce higher error scores than in the massed condition. Whatever structural information can be recovered from the cue, however, will tend to mitigate differences in error scores between the two spaced conditions. For that reason, I expect minimal difference in error scores at the two spaced conditions.

2. Previous Research

This review of the literature situates the study of spacing effects in singers memory within the larger context of memory research. It includes a review of human memory processes involved in the coding and recall of information, both non-musical and musical. In addition, the effect of distributed learning on memory retention is examined, as are the particular characteristics of memory for song.

2.1 Stages of Learning and Memory

It is a well-established principle of cognitive psychology (the study of mind) that what we call 'memory' is actually several different systems. Atkinson and Shiffrin (1968) proposed the existence of three different memory stores, the *sensory register*, the *short-term store* and the *long-term store* (Crowder, 1976/2014). Information flows from the environment into a sensory register, an extremely short-term memory store relating to particular sensory input systems. Of these, the most studied are *iconic* memory, the store for visual sense impressions lasting on the order of about 500 milliseconds; and *echoic* memory, the system storing very short-term auditory memory, averaging about three seconds in length (Baddeley, 2004). The sensory traces that are then attended to, identified and recoded enter into the short-term store (also called short term memory, or STM), a limited capacity system capable of holding about 5-7 items (Miller, 1956). If this information is not maintained in STM through conscious attention, it is lost. In order to assure retention beyond the limited span of STM, where decay rates are relatively rapid, information must be transferred to the long-term store (also known as long term memory, or LTM). This transfer of information to LTM (an almost permanent store of memory) is the learning or acquisition process (Crowder, 1976/2014). This, together with retention or storage, whereby memory persists over passing time, allows for the retrieval of information from storage when it is needed.

LTM is a multi-part system, fractionable into *explicit* (declarative) or *implicit* (procedural) memory (Baddeley, 2004). *Declarative* memory is a store of knowledge **about** things. It is knowledge that can be put into words. Declarative memory can itself be divided into two separate systems, *semantic* and *episodic* memory (Tulving, 2002). Semantic memory is an abstraction or generalization from the original perceptual data of experience: it consists largely of facts rather than episodes. The second type of explicit memory, episodic memory, allows us to review past events we have experienced. Through episodic memory we can "recollect specific incidents from the past, remembering incidental detail that allows us in a sense to relive the event…" (Baddeley, 2004, p. 5). Episodic memory is considered to be linked to autobiographical memory, the ongoing story of the self, or 'self-memory system' (SMS) (Conway & Pleydell-Pearce, 2000). Episodic information is knowledge contained in episodic memory and is largely sensory-perceptual in nature (Conway, 2001).

> We conceive of episodic memory as a system that contains experience-near, highly event-specific, sensory-perceptual details of recent experiences…These sensory-perceptual episodic memories do not endure in memory unless they become linked to more permanent autobiographical memory structures. (Conway, 2001, pp. 1375-1376)

Concurrent thoughts, emotional condition and body-state may all be part of the contextual information specifying a particular experience.

Baddeley has this to say about the relation of semantic to episodic memory: "While it is generally accepted that both semantic and episodic memory comprise explicit memory, as opposed to implicit memory systems, the relationship between the two remains controversial."

(Baddeley, 2004, p. 5). While some studies have been conducted relating semantic memory to retrieval strategies in music performance (Segalowitz, 2001; Chaffin & Imreh, 2002; Ginsborg & Chaffin, 2011) which I will examine in more detail below, accessing episodic memory in performance ('reliving the event'), is less well understood.

The other side of LTM from explicit memory is implicit or *procedural* memory. Procedural knowledge is the knowledge of how to do things. It guides skills like chess-playing or music-making, skills which are difficult to express verbally (Ten Berge & Van Hezewijk, 1999). It is memory that is accessed through performance, rather than through recollection (Baddeley, 2004). The clearest evidence for the separability of procedural from declarative memory is shown by amnesiacs (Warrington & Weiskrantz, 1968). Densely amnesiac patients show enhanced identification of previously presented items, although they are not aware of having seen them before. In certain cases of severe amnesia the ability both to play and recognize music is preserved even though brain areas associated with the formation and retrieval of LTM have been almost totally destroyed (Finke, Esfahani & Ploner, 2012). Learning of a wide range of types can be preserved as long as learning does not require retrieval of a particular episode.

The implicit memory system (IMS) has another important form, *implicit perceptual representation*. It is unconscious statistical learning that "keeps a record of regularities in the environment and structures unconscious expectation about environmental events" (Snyder, 2008, p. 108). This may have bearing on singers' memory. Where explicit structural analysis may play an important part in memory strategies for instrumental musicians (Chaffin & Imreh, 2002; Chaffin, Logan & Begosh, 2008) and some singers (Ginsborg, 2004; Ginsborg & Chaffin, 2011) other singers memorize effectively with no conscious attention to structure (Rubin, 1995). An implicit understanding of musical structure could help to explain memory for song in cases where there seems to be no overt structural analysis.

Working memory. The crucial question for any study of learning and memory is how information is transferred from short-term to long-term memory. To account for this, Baddeley & Hitch (1974) introduced the idea of STM as *working memory*, a multi-part system. At the centre of working memory is the *supervisory attentional system* or SAS, assisted by two subsidiary systems, the *phonological loop* and the *visuospatial sketchpad* (Baddeley, 2004). The phonological (or articulatory) loop is assumed to comprise a store that holds memory traces for a couple of seconds, combined with a subvocal rehearsal process. This is capable of maintaining the items in memory using subvocal speech, which can also be used to convert nameable but visually presented stimuli, such as letters or syllables, into an aural or *phonological* code. Baddeley also mentions a related system, that of *auditory imagery*, as a distinct subcomponent of working memory.

The visuo-spatial sketchpad functions analogously to the phonological loop. It allows the temporary storage and manipulation of visual and spatial information. A fourth component of WM, the episodic buffer, provides a “multimodal temporary store of limited capacity that is capable of integrating information from the subsidiary systems with that of LTM.” (Baddeley, 2004, p. 4). Baddeley’s revised version of working memory (Baddeley, 2004) can be regarded as a fractionation of the original model into separate attentional and storage systems, with the fluid systems being attentional and the crystallized systems representing storage (Figure 1). Specialized systems of language and of visual semantics are added in order to illustrate the integration of these brain functions into working memory without the necessity of conscious attentional control. “I see WM as a complex interactive system that is able to provide an interface between cognition and action, an interface that is capable of handling information in a range of modalities and stages of processing." (Baddeley 2012, p. 18)

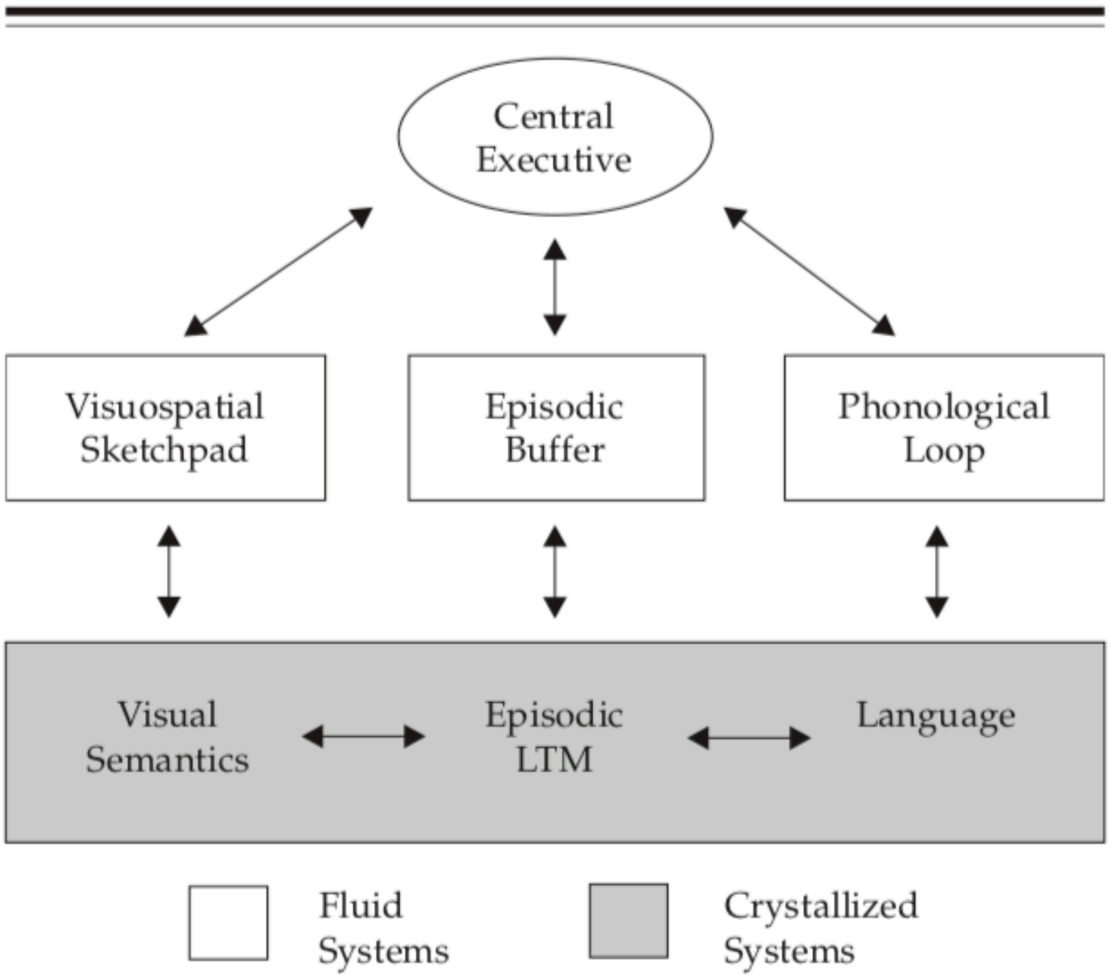

Figure 1. Baddeley's expanded model of working memory, revised to incorporate links with long-term memory (LTM) by way of both the subsystems and the episodic buffer (Baddeley, 2000).

The episodic buffer is also responsible for the *chunking* of information in STM. Chunking allows for the storage of groups of elements rather than single elements (Miller, 1956). Where STM holds only 5-7 single elements, chunking allows for the recall of 5-7 groups of elements. For example, separating a phone number into an area code, then a three-digit, then a four-digit number transforms 10 digits into three chunks. The structure of the chunk is determined by a *schema* or pattern that creates associative relations between the elements. Schema, a concept first introduced by Bartlett (1932), is said to be the essential organizing system of the brain for storage and retrieval of knowledge (Bower, 2000): “Schemas represent knowledge about some domain and capture it using clusters of organized expectations” (Bower 2000, p. 24). Schemas may be of any size or level of detail and may contain other schemas nested within them. For example, “face” contains schemas for “nose”, “mouth”, “eyes” and many other elements. Common schemas may include a multitude of possibilities, including objects, ethnic and personality stereotypes or routine actions. New schemas are acquired through experience; new knowledge is fit into old schemas as a way of understanding events. Structured schemas can also be viewed as large clusters of elementary associations (Bower, 2000). Once novel information has been fit into a pre-existing schema, inferences can be drawn about the new information based on these associations. In musical terms, creating schematic associations through chunking allows notes to become motifs, motifs to form phrases or periods, and periods to form sections. Musical forms (sonata or partita for example) may be thought of as different kinds of overriding schema.

In response to a growing body of experimental evidence, many other theories of working memory have been formulated since the introduction of Baddeley’s original model of working memory in 1974. Miyake and Shah (1999) examined 10 different theories of working memory, including Baddeley. In general, they represent an increasing fractionation of working memory into various distinct subsystems. There is also evidence that Baddeley’s working memory model may not be enough to account for musical memory. A specialized musical phonological loop in working memory may exist (Berz, 1995). “Differing degrees of disruption in a variety of experimental memory storage tasks also supports the view that musical information is held in a different area of STM than is verbal information." (Berz, 1995, p. 360). For example, Martin, Wogalter and Forlano (1988) found that unattended speech or song was more disruptive of reading comprehension than unattended instrumental music. On a music reading/identification task, unattended instrumental music had a greater disrupting effect than unattended speech. The conclusions were that disruption to attention occurs in a modality specific form, implying differential processing for music and for speech in working memory, and a similar processing mechanism for speech and for song. The final experiments in the group established that the interference from spoken text on reading comprehension was a matter of semantic rather than `phonological distraction.

There is evidence for a somato-sensory perception and production network for singing distinct from that of speech (see Zarate, 2013 for a review). Patel (2010) makes a compelling case for overlap between syntactic processing of music and language, and dissociation of syntactic representation of music and language (Patel 2010, pp. 276-298). The degree to which song occupies its own mode of processing in working memory distinct from speech and instrumental music remains an open question.

The basic systems model of episodic memory. The problem with Baddeley's working memory model is that, while it works well for relatively simple stimuli that are "unimodal" (e.g. words paired in a list), it does not account well for "complex, real-world situations involving multiple senses, language, narrative and emotion" (Rubin, 2006, p. 279). The basic systems model proposes that memory depends "on the unique functions and properties" of a series of basic systems (Rubin, 2006, p. 277). Each system contains its own knowledge which forms part of the memory of the event. According to the theory, all declarative memory, both episodic and semantic, is constructed from the interactions between 11 basic systems; vision, audition, olfaction, other senses, spatial imagery, language, emotion, narrative, motor output, explicit memory and search and retrieval (Rubin, 2006, p. 279).

> Each system has its own functions, neural substrate, processes, structures, kinds of schema, and types of errors that have been studied individually. Each also has its own forms of memory, possibly including its own sensory information store, working memory buffer, and long-term memory. (Rubin, 2006, p. 277)

In the basic systems model, there are three types of coordinating systems for memory retrieval. The first is an explicit retrieval system "that binds together everything that occurred at the same time into one event that could later be retrieved as a memory" (Rubin, 2006, p. 286). This is the explicit memory system. The second system modulates the encoding of memories on the basis of surprise, interest or emotional arousal. This is handled by the emotion system. A third system searches "for some components of a memory when cued by other components, while using inhibitory mechanisms to suppress dominant responses that do not fit the criteria set by the known cues." (Rubin, 2006, p. 286). This is primarily a function of the search-and-retrieval system.

The basic systems model is derived from a study of complex behaviour, the cognitive profile of memory for ballads, epics and children's rhymes (Rubin, 1995). It justifies the consideration of systems other than the semantic structures of explicit memory in the formation of complex memories, and especially, memory for song derived from the oral tradition.

Encoding and retrieval. The nature of coding processes (i.e. whether meaningful and highly associated or not) has an important influence on the likelihood of retrieval (Bellezza, 1996). The levels of processing theory (Craik & Lockhart, 1972) sought to explain the memory advantage for stimulus materials that are meaningfully connected to a retrieval cue as being the result of deeper processing. Reformulations of the theory emphasize breadth of association rather than depth of association as an explanation of the effect (Roediger, 1993). The nature of the retrieval cue is also a powerful influence on the likelihood of retrieving a memory (Tulving & Thomson, 1973). Tulving's *encoding specificity* principle states that "a retrieval cue will be effective to the extent that information in the cue was incorporated in the trace of the target event at the time of its original encoding" (Brown & Craik, 2000, p. 99). Tulving also found that generating other possible candidates for recall is unnecessary when the original contextual elements are supplied. This means that the mind does not have to review alternative material to come up with the target content, provided the contextual cues are strongly and exclusively associated with the original material. This has ramifications for complex materials like narrative prose, poetry, music, and in particular, song, which offer ongoing contextual cues embedded in the material itself. Finally, state-dependency, the person's mental or physical state during encoding and retrieval, can also has an effect on memory. Ideally, state at retrieval should match the state at encoding for maximum retrieval effect (Brown & Craik, 2000).

The new theory of disuse. Human memory is not a literal recording, but a storage system based on what we already know. New items are fit in to memory in terms of their meaning. According to Bjork and Bjork (1992), that meant in terms of their semantic relationship to items, schemas, and scripts already in LTM. The act of storing new information in memory appears to create the opportunities for additional storage. The more knowledge we have in a given domain, the more ways there are to store additional information (Bjork & Bjork, 1992). Once entrenched in LTM, material remains in memory for an indefinitely long period of time. When we retrieve information, that information becomes more retrievable, and other information can become less retrievable. This argues against a purely associative theory of memory (Bjork &

Bjork, 2007). The more difficult a successful retrieval, the more it promotes learning. One final peculiarity of the human memory system: memory representations constructed earlier tend to become more accessible over time, relative to later constructions, a matter of great significance to performers. This regression process helps to determine which items are retrievable from memory storage.

On the basis of these and other characteristics, Bjork and Bjork formulated a "new theory of disuse" to explain memory functioning. According to the theory, an item in memory is characterized by its storage strength and its retrieval strength. The likelihood of its being remembered in a given moment is wholly a function of its retrieval strength in response to a given cue. Since retrieval strength is increased when the storage strength of the item is greater, how strongly associated the item is in memory with the cue material is still of importance.

Storage strength grows through learning, that is, as a result of opportunities to rehearse and retrieve the item. As storage strength increases, however, the effect of learning decreases: it becomes more difficult to add to the storage strength of the item. Two hours of study of an item will not give you twice as much learning (increase in storage strength) as one hour. In addition, high retrieval strength will retard the increase in storage strength. That is why an easily remembered item (high in retrieval strength) will not add as much to learning (a gain in storage strength) as remembering an item which is more difficult to retrieve. This offers an elegant justification for many separate memory trials as a preparation for a memorized performance. Each difficult memory retrieval of a piece of music may add to the chance of retrieving that material at another time, as long as there is an opportunity to identify and correct the errors.

2.2 Memory for Music

Music performance relies on a continuously unfolding chain of associative elements (Chaffin, Logan & Begosh, 2008). It is a fluent retrieval of effectively coded musical material. Almost all the contents of LTM are not conscious at a given moment although situational associations may cue memories into consciousness. “Memories that form spontaneously while learning a new piece take the form of associational chains in which each passage cues the memory of what comes next.” (Chaffin et al., 2008, p. 352). Each of the separate perceptual systems involved in performance lay down their own trace in episodic memory (Rubin, 2006). Music is an affective language, as well as a highly organized intellectual structure. Information from auditory, motor (or kinesthetic), visual, emotional, narrative and linguistic memory may all interact in the memorized performance (Chaffin et al., 2008). “In performance, auditory memory tells the musician what comes next, providing cues to elicit the music from memory, while also letting the musician know that things are on track”. Motor memory is fundamentally procedural, and “provides the clearest examples of associative chaining in memory: each action in the series cues the next” (Chaffin et al. 2008, p. 355). Instrument-specific context has an effect on music performance. Mishra and Backlin (2007) found that changing instruments between learning and practice led to context-dependent environmental effects in performance of a short memorized piano work. Changing physical locations between learning and performance, however, had no effect. The authors suggested that, following Glenberg (1997), associative cues in musical material take precedence over cues related to environmental context, unless that context is also implicated in the learning strategies used. In this case, the proprioceptive cues offered by the particular piano used may have been important in both coding and retrieval.

In a study applying the levels of processing theory to piano music memorization, Segalowitz, Cohen, Chan and Prieur (2001) taught a short piano work to 16 pianists. Participants were exposed to four different learning combinations, high and low processing and high and low elaboration. Deep processing was equated with aesthetic judgment, shallow processing with attention to surface characteristics of the music. High and low elaboration were expressed through amount of practice in each condition. Although the experiment was seriously underpowered, memory for the music was more accurate in the high elaboration, deep processing condition, and least accurate for the shallow processing, high elaboration condition. Problems with the original levels of processing theory were dealt with by explaining the effect of the experiment according to a transfer-appropriate model (Lockhart, 2002), whereby coding strategies used in learning are effective as cues when they are present at the time of retrieval. The authors suggested that non-aesthetic elements (resulting in shallow processing) are inappropriate for transfer to the recall phase of the experiment. The implication is that in retrieval, musicians are guided by aesthetic qualities.

Performance cues, a type of declarative, semantic-based cue specific to expert music performance, allow memory retrieval at pre-determined points in the musical structure. A study of the concert pianist, Gabriela Imreh, preparing the third movement of the Italian Concerto of Bach for memorized performance showed her reliance on performance cues to enable the fluent memory retrieval of an extremely difficult piece of keyboard music (Chaffin & Imreh, 2002). The third movement of Bach's Italian Concerto has no breaks in the ongoing rhythmic pattern, a very quick tempo, and many parallelisms in the music, points where a small change in the ongoing figuration leads to a variable outcome. Memory security in this type of writing may necessitate content-addressable cues so that performance can be resumed at the cue location in the event of a memory slip. Their applicability to vocal music has not been convincingly demonstrated.

In a longitudinal study of memory processes in vocal performance, Ginsborg, a cognitive psychologist and singer, observed herself preparing the Ricercar 1 from the Stravinsky Cantata (Ginsborg & Chaffin, 2011). Ginsborg and her conductor notated the scores after the performance to establish their understanding of the formal structure of the piece, and to establish the musical "landmarks" used as individual and as shared performance cues. The study may suffer from the inherent bias in a cognitive psychologist studying her own performance, and the necessarily semantic procedure involved in verbal discussion of musical process. It is interesting that the author differentiates between those structural cues used in rehearsal and for subsequent written recall (landmarks) and those procedurally-based cues relating to physical-motor ability used in performance (triggers). Unlike Chaffin's study of the pianist Gabriel Imreh (Chaffin & Imreh, 2002), this study does not demonstrate the use of semantically-based declarative cues in performance. Song in the oral tradition, moreover, is not content-addressable (Rubin, 2006). That means a memory slip can only be corrected by returning to the sectional break in the material, and conscious, declarative memory cues are not being used.

There is strong evidence that memory for melody is based on tonal structure and melodic contour (Schmuckler, 2016; Dowling, 1978). Melodies in Western tonal music imply an underlying harmony (Patel, 2010). Transposed melodies are recognized more accurately when the implied harmony is based on the simplest tonal formulas (Cuddy, Cohen & Mewhort, 1981). Simple melodies are best suited for optimal syllable retention (Wallace, 1994; Rainey & Larsen, 2002; Purnell-Webb & Speelman, 2007). Melodies using pitches that outline tonic triads at phrase endings are remembered more accurately than comparable melodies without, but only when those triads are employed within regular accent patterns (Boltz, 1991). Modal melodies can be difficult to learn compared to melodies based on conventional tonic-dominant harmonic progressions (Oura & Hatano, 1988). Words set to melody with an irregular phrase structure may be more difficult to remember than a simple recitation of text (Racette & Peretz, 2006). In general, simpler melodies are more likely to facilitate the retrieval of words in performance.

2.3 Memory for Prose

Among the earliest studies of meaningful prose were those by Bartlett (1932). Bartlett used observation derived from a real-life approach which was closer to naturalistic behaviour than the usual laboratory models. Although his observations lacked a scientific hypothesis-driven framework and any attempt at statistical analysis (Paul, 1959), the fundamental theoretical framework he formulated for understanding memory of discourse is still relevant. In a series of studies using “The War of the Ghosts”, a story of the Bella Coola people transcribed by Boas (Boas, 1898/2006), Bartlett found that memory for longer prose texts was schematic and based on reconstruction, not on verbatim recall. “Although the general form, or scheme of a prose passage thus persisted with relatively little change, once the reproduction had been effected…. the actual style of the original was nearly always rapidly and unwittingly transformed.” (Bartlett, 1932, p. 83). Bartlett defined schema as “an active organization of past reactions, or of past experiences, which must always be supposed to be operating in any well-adapted organic response” (Bartlett, 1932, p. 201). Schemas allowed the organization of local elements into an overriding whole. Detail is retained when it fits with pre-formed interests and tendencies. The process of rationalization or reconstruction, often based on an affective attitude, “gives the whole that specific ground, frame or setting without which it will not be persistently remembered”. (Bartlett, 1932, p. 94). “Attitude is very largely a matter of feeling or affect… The recall is a construction made largely on the basis of this attitude, and its general effect is that of a justification of the attitude" (Bartlett, 1932, p. 207). Whereas memory for detail was very weak, memory for the story outline was relatively stable and could be recalled months or years later.

Cofer (1941) compared verbatim and gist learning for two different narrative prose passages adapted, like Bartlett, from transcriptions by Boas of traditional tales of the Bella Coola and Kwakiutl peoples (Boas, 1898/2006). For verbatim learning, criterion was reached in a single learning session, demonstrated by two perfect recitations. He found that verbatim learning for narrative prose was much more difficult than logical learning, especially for longer passages. Retention at nine months could not be evaluated statistically due to the small number of participants. Using specially written scientific stories, Sachs (1967) examined recognition memory for sentences embedded within longer sections of prose. Subjects retained memory of semantic changes, but not of syntactic changes. Only immediate or short retention intervals filled by intervening prose were examined. Meyer & McConkie (1973) looked at recall for two different prose passages, 481 and 502 words long. Testing immediate recall in written form, they found that words in prose passages were recalled according to their position in a logical structure.

For Kintsch and van Dijk, "memory for verbal utterance is predominantly semantic" (Kintsch & van Dijk, 1978, p. 61) and verbatim memory for prose is poor after even a short retention interval. "Discourse processing (understanding, organization, retrieval) is a function of the structures assigned to the discourse during input." (Kintsch & van Dijk, 1978, p. 66). These structures are largely a matter of semantic representation. Narrative structures define the characteristics and particular cognitive processing of stories. They may include such things as setting, complication, resolution, evaluation, and moral and define the hierarchical syntax of narrative (Kintsch & van Dijk, 1978, p. 69). Comprehension and recall of discourse (and of semantic information in general) is also based on relations between the discourse and implied elements coming from particular or general schemas representing our existing knowledge, interests and intentions.

Noice and Noice (1997) offer a very different view of discourse processes from Kintsch and van Dijk. In studies of actors' memory, they found that actors are able to reproduce long sections of text verbatim through two approaches. The first is through use of a coherent coding system, similar to Kintsch and van Dijk's "structures assigned to the discourse during input." (Kintsch & van Dijk, 1978, p. 66). Actors use depth analysis of the material through elaboration, depth of processing, distinctiveness, causal attribution, perspective taking and overlearning (Noice & Noice, 2002, p. 11). This means that the particular choice of word in any given context is completely constrained by the actor's understanding of the mental, emotional and circumstantial environment that produced those words. In effect, no other words are possible for the speaker of those words at that moment. The second approach is "active experiencing" or actually doing and feeling what the character is doing or feeling (Noice & Noice, 2002, p. 10). It is not "purely cognitive, but necessarily involves the type of cognitive-emotive-motor processing that would be inherent in any human interaction" (Noice & Noice, 2002, p. 10). As it turned out, however, the crucial element was the active involvement of that "cognitive-emotive-motor system". Generally speaking, the higher the emotionality of the participant conveying the text, the greater the effect. Noice and Noice offer a compelling experimental verification for Bartlett's intuition that the emotional "attitude" is of prime importance in the long-range verbatim retention of discourse.

2.4 Memory for Poetry

Although I could find no recent laboratory-controlled studies on memory for verses of poetry over longer retention intervals, there is a fair amount of research indicating that poetry offers a memory advantage over a comparable prose setting. Ebbinghaus (1885/1913) found that learning six stanzas of poetry took, on average, one-tenth the time of learning a comparable number of nonsense syllables (Ebbinghaus, 1885/1913, p. 51). Using poetic and rhetorical[3] materials, Rubin (1977) offered a very different view of long-term verbatim recall from the prose studies listed above. Rubin found that university undergraduates remember long stretches of five familiar texts ("The Preamble to the Constitution," "The 23rd Psalm: A Psalm of David," "Hamlet's Soliloquy", “The Gettysburg Address “and “The Star-spangled Banner”) verbatim, through associative chaining of surface elements. They showed no evidence of remembering in an abstractive, reconstructive manner. Furthermore, recall was accurate and organized in terms of surface structure units.

The prose materials in Rubin’s study (the Gettysburg Address and the Preamble to the Constitution) share certain features with poetry, which could help to account for their memorability. They are rhetorical pieces written to be delivered in public address. They all have rhythmic patterning (not always regular), alliterative devices and phrasing divided by points to take breath. They were often learned by memory through frequent exposure in early life and all have important emotional resonance for American students, offering a significant “attitude” (to use Bartlett’s terms). The prose materials eliciting little or no verbatim recall mentioned above included non-poetic (factual) translations of traditional stories (Bartlett, 1932; Cofer, 1941), passages from Scientific American (Myer & McConkie, 1972), or a 1500-word story from the

[3] By this I mean written with the devices of rhetoric and intended for public address.

Decameron (Kintsch & van Dijk, 1979), too long to be comparable to other verbatim learning studies. Generally speaking, these are materials without poetic or alliterative devices, and without the rhythmic characteristics of rhetorical or poetic text.

While recognition memory for specific words in prose separated by intervening text diminishes greatly over short retention intervals (Sachs, 1967), recall of phrases in lyric poetry is not diminished (Tillman & Dowling, 2007). Moreover, verbatim memory for surface features of target syllables in poetry is better than for target syllables in prose. The authors suggest that both music and poetry offer semantic structures that facilitate recall of surface features based on rhythmic structure and temporal organization. Similar alliterative lines of poetry are more likely to be falsely recognized in both immediate recall and after 12 hours than non-alliterative lines or paraphrases, indicating that alliteration as a formal, schematic device is preserved in memory and helps to cue memory (Atchley & Hare, 2013). Alliterative cues reactivate memory of previous information that is phonologically similar, effects holding for both poetry and prose (Lea, Rapp, Elfenbein, Mitchel & Romine, 2008). A continuous reading paradigm was used, so the effect of retention interval was not tested. Undergraduates will select words to complete sentences based both on rhyme and on meaning, supporting the importance of surface features in determining word choice (Rapp & Samuel, 2002).

2.5 Memory for Song

Most of the research into song memory has focused on song as a mnemonic, a device for remembering text. As such, it usually compares sung and spoken versions of the same text as heard by the learner. Recall or recognition of the text is then evaluated at short retention intervals. Responses are frequently in written form. Many (though not all) studies of the effect of hearing sung materials on the retention of text show advantages to sung over spoken presentation. This may be due as much to the rhythmic nature of the setting as to the influence of the melody.

Wallace (1994) in a series of experiments, examined the degree to which recall of words was influenced by the presence or absence of a melodic setting. The first experiment compared five sung or spoken heard repetitions of the same three verses excerpted from a traditional ballad among a population of undergraduates. Subjects were instructed to transcribe the text verbatim after the first, second and fifth repetitions. Results indicated the superiority of the sung over the spoken version for verbatim retention of the words at immediate testing and testing after a retention interval of twenty minutes. Testing after the second repetition of the text also indicated that participants hearing the sung version were likely to rely on structural characteristics of the rhythm to indicate the correct number of syllables, even when they could not remember the actual words. They also used the correct line breaks between lines of the text to facilitate word retrieval. A second experiment used a synchronized metronome beat with the spoken condition and found no memory advantage to rhythmic speech over the other conditions. In a third experiment, when exposed to a single verse instead of three verses, better verbatim recall was found for the spoken condition over the sung condition. The author concluded that increased familiarity with the melody in the three-verse condition led to greater sung over spoken retention. Finally, when three different melodies for the same lyrics were compared to a spoken version and a version with a single (original) melody, the melodic-repeated condition was found to be superior to the spoken and three-melody versions for verbatim recall. In sum, verbatim recall was found to be highest for a sung condition, but only where the melody was sufficiently or easily learned. Providing strong rhythmic intonation and beat in spoken presentation of the text did not aid text retrieval, nor did providing three different melodies for the same text.

Purnell-Webb and Speelman (2008) demonstrated in a careful re-evaluation of Wallace (1994) that rhythm was as effective a mnemonic device as a familiar melody, provided there was sufficient rhythmic information available. They used a pattern of strong and weak beats accompanying the text rather than Wallace's undifferentiated metronome beat. Increasing familiarity with a melody did not influence recall. Purnell-Webb and Speelman also had the participants write down the syllables at every hearing. Under these conditions, it is not possible to know what part of the processing of verbal materials from short term to long-term memory was played by the aural exposure to the stimulus, and what part by the act of writing down the materials.

Chazin and Neuschatz (1990) found a mnemonic effect for free recall of a list of minerals set to a familiar song ("Mary had a little lamb"), but only in an immediate memory test. At a recall interval of one week, no difference was observed between the sung version and the spoken lecture. Only three listening opportunities in one session were given to participants, so it is possible that the initial learning was insufficient to show an effect on longer-term recall. Examination of the lyrics used also shows poor agreement between the rhythmic structure of the text and the accent pattern of the melody, a variable of known importance for memory of words set to music (Serafine, Davidson, Crowder & Repp, 1986).

McElhinney and Annett (1996) demonstrated that those who learned a passage through song remembered more of the text and were more likely to chunk or group the material as mnemonic support than those who learned through speech. Prior exposure to the stimulus materials (a song by Billy Joel) by some participants cannot be entirely ruled out. Hearing the Prelude to the Constitution in a sung condition resulted in verbatim recall superior to the spoken condition (Calvert & Tart, 1993). Undergraduates heard the material in eight sessions spread over four weeks, as excerpts produced in a video format. Although immediate recall after the eight sessions produced stronger results for the sung condition over the spoken, the discrepancy between sung and spoken conditions was even greater after a retention interval of five weeks. Because there was no aural-only condition, the contribution of the visual aids presented in the programs cannot be assessed independently of the aural presentation. The study was also underpowered, with only seven participants in each of four experimental conditions. In addition, prior knowledge of the stimulus materials may have played a differential role in recall across conditions.

Calvert and Billingsley (1998), in the first of two experiments, taught English-speaking pre-school children via video presentation to sing “Frère Jacques” in an English (comprehensible) or French (incomprehensible) version, with a verbal recitation at testing. Testing after one session or four daily sessions occurred immediately after the final presentation. The children remembered the incomprehensible French lyrics sung to a simple tune better than words in their own language (English). In a second experiment children were asked to learn and remember their phone number through sung or spoken presentations. Results after six sessions indicated higher results for the spoken condition, and uniformly poor results for comprehension. The authors speculated that for young children there was competition between the meaning of the words and cognitive requirements of learning to sing the song: learning to sing the song actually inhibited semantic processing of the words. However, the authors used a novel tune for the song condition, which may have led to greater difficulty with processing and consequently poorer results.

There were no initial advantages to hearing names of baseball players set to the tune of “Pop goes the Weasel” over a spoken version for undergraduates. However, at a one-week retention interval, participants consistently showed fewer trials to reach the criterion of a memorized performance (Rainey & Larsen, 2002). The authors speculated that melodic setting could improve immediate recall for participants (above the pre-school age) for meaningfully-connected textual materials, as in Wallace and Rubin (1988b), Wallace (1994), and McElhinney and Annett (1996), but not for unconnected materials, as in this study or in Calvert and Billingsley (1998), experiment 2. Murakami (2017) used a 16-word grocery list set to a newly-composed four-line song to test a speech, rhythmic speech, melody only or melody with harmony condition in healthy older adults. Results showed recall in the rhythmic speech condition to be significantly more accurate than the other conditions at immediate post-training testing. At a retention interval of 10 minutes there was no significant difference between conditions. It would be interesting to know if a follow-up test at a longer retention interval would show a differential effect on memory for the learning conditions. Setting nonsense text to musical pitch produces higher recognition scores than spoken syllables alone for adults (Schön et al., 2008). The authors concluded that the musical setting reinforced structural boundaries in the text through enhancing levels of arousal and attention and by using changes of pitch to enhance phonological discrimination.

Equating the presentation rates of the material resulted in no appreciable advantage to sung over spoken text retention for undergraduates (Kilgour, Jakobson & Cuddy, 2000). A comparison of the songs used as stimulus materials in the Wallace (1994) and those used in Kilgour et al. (2000) does reveal some differences, however. The songs used by Wallace are traditional ballads with four phrases and no chromatic tones or modulations. One of the songs used by Kilgour et al., “Out in the Moonlight” by Dougherty, has eight phrases, with phrases 1-2 and 5-6 the same. Phrases 3-4 and 7-8, however are not the same. In addition, there are chromatic tones in phrases four, seven and eight and an implied modulation in phrase three. “Willcott”, a song by Marshall, also used by Kilgour et al., has a four-line structure more similar to the Wallace materials. However, unlike the tunes used in Wallace, the melody for the third phrase is built around the sixth scale degree, arguably adding extra complexity. The results in Kilgour et al. would have been more convincing if they had consistently used stimulus materials more similar (or identical) to those used in Wallace (1994).

Other negative findings for melody as a mnemonic device may also be explained by the particular characteristics of the stimulus materials. Racette and Peretz (2007) found an advantage to memory for lyrical text in the heard/spoken over the heard/sung condition. They used a group of newly composed tunes by Gagnon that lacked a distinctive melodic profile and tested immediately after the end of the learning session. Using the same materials, a study of the effects of a vocal setting in memory for lyrics compared a small group of participants with Alzheimer’s dementia in a variety of learning conditions (Moussard, Bigand, Belleville & Peretz, 2014). They found a retention advantage in sung lyrics set to a highly familiar tune over speaking lyrics, singing lyrics set to a well-learned unfamiliar tune, and hearing spoken lyrics for the normal control group when tested at a ten-minute retention interval. As with the Racette and Peretz (2007) study, they did not test to criterion in the first session and had no subsequent training session for the normal controls, limiting its usefulness as a memory study. In addition, both studies taught participants a group of six different songs with different lyrics on six different occasions, leaving open the possibility of proactive interference affecting results.

Most of the studies cited above use a short (immediate or ten-minute) retention interval for testing. Two of the studies used longer retention intervals, one week for Rainey and Larsen, (2002) and five weeks for Calvert and Tart (1993). Both of these studies showed evidence for greater recall or words when listening to song over listening to speech. Results at short RIs (10-20 minutes) are equivocal. While Kilgour, Jakobson and Cuddy (2000) present a legitimate objection to Wallace‘s (1994) procedure for tempo equalization of stimulus materials, their own musical stimulus presented difficulties not found in Wallace’s material. Purnell-Webb and Speelman (2007) were more convincing as a refutation of Wallace, using the melody to "Scarborough Fair" itself and in recomposed versions which corresponded well to the structure of Wallace's stimulus materials. They showed no memory advantage at a 15-minute retention interval for materials heard as song compared to materials heard in a rhythmic recitation. The results for Murakami (2017) also supported Purnell-Webb and Speelman in finding no memory advantage for hearing a sung version of a grocery list over rhythmic speech or spoken recitation at a ten-minute retention interval. In general, the benefits of song as a mnemonic are modulated by 1) the melody's familiarity and simplicity, 2) the amount of repetition in the learning phase, 3) the length of the retention interval and 4) the poetic qualities of the text.

There are very few studies which have taught participants to sing the text, and then evaluated retention of the text in a sung-back form at testing. Among the most important are Ginsborg (2002), Ginsborg and Sloboda (2007) and Good, Russo and Sullivan, (2014). The Ginsborg (2002) study was an observational study conducted remotely in subjects' homes. A learning rather than a memory study, it used the number of learning trials and accuracy at the final learning trial to assess three different approaches to learning an unfamiliar two-verse song. Ginsborg and Sloboda (2007), a controlled study of different coding conditions in song learning, used a single 10-minute RI. They found that singers with higher expertise benefit from studying words and music together. The Good, Russo and Sullivan (2014) study of sung materials in language learning taught 38 Spanish-speaking children a four-line lyrical passage in a sung or spoken English version. Participants were trained using four learning sessions of 20 minutes each, spread out over two weeks. Testing was immediate, at one day, and at a retention interval of 6 months. They found higher scores for pronunciation (after the third session), recall (after the fourth session) and translation (one-day later). At six months, scores for pronunciation and recall (but not translation) were higher for the sung condition compared to the spoken. It is interesting that in this study, the translation of the sung text was given in a handout, and in a paired associate recitation after the sung portion of the session. It is not clear to what degree an actually sung translation in the learning session would benefit the participants.

Memory for ballads. Wallace and Rubin (1991) analyzed four groups of songs drawn from the Brown collection of traditional North Carolina ballads, edited by White (White, 1952-1962, Vol. 2). They found certain commonalities, including: (a) four-line verses with no chorus; (b) end rhyme at the ends of lines two and four with occasional assonance, alliteration, and rhyme within the lines; (c) four feet per line where a foot consists of an unstressed followed by a stressed syllable; (d) a vocabulary consisting mostly of one-syllable words; (e) meaning carried predominantly by the last two lines in the verse, whereas imagery was carried predominantly by the first three lines of the verse; (f) approximately 20% of the lines were echoes of other lines within the ballad and 5% were lines used in other ballads, and (g) strong agreement of musical and textual stresses (Wallace & Rubin, 1991, p. 199). In a subsequent study, the authors conclude that ballads “are a highly structured stimulus domain with many different characteristics that play individual and interacting roles in making ballads easy for those who are experts in the genre to learn and remember” (Rubin, Wallace & Houston, 1993, p. 437).

Wallace and Rubin (1988b) also examined constraints within ballads for their effect on recall in a population of non-specialists. Twenty-seven undergraduates listened to ten repetitions of an unfamiliar ballad and were tested for word retention (in writing) after ten minutes. Imagery, metrical agreement and causal connectedness all correlated significantly with recall, all features that had been observed in expert ballad performers (Wallace & Rubin, 1988a). Furthermore, when pairs of words in the same ballad were changed so that instances of assonance and alliteration were removed, significantly fewer of the changed words were recalled. Participants were "more likely to recall the designated word pair, heard within the verse, verbatim when the poetic constraints were present than when those constraints were missing. In the nonpoetic case, subjects recalled the gist of the word pairs but not the verbatim word pair" (Wallace & Rubin, 1991, p. 183). Finally, where spoken recitation was heard, those lines which corresponded most closely to the overall metrical pattern were remembered best in a rhythmic recitation, a result consistent with rhythmic information acting to cue word recall. (Wallace & Rubin, 1988b). "Thus, even for subjects who are not familiar with the tradition, the presence of some characteristics can improve verbatim recall" (Rubin, 1991, p.183). The different constraints can be regarded as schemas, not just for meaning, but for poetics, rhythm, imagery and music (Wallace & Rubin, 1988b).

In an experiment testing 127 undergraduates for the effectiveness of rhyme and meaning used individually and then together as cues, Rubin and Wallace found that the probabilities of responding with the target words, given the rhyme, meaning, and dual cues, were .192, .142, and .973, respectively (Rubin & Wallace 1989, p. 703). The observed effect for dual cuing was three times the maximum predicted under existing models. A specific example taken from Rubin and Wallace (1989) is illuminating. The linguistic/semantic cue "building material", for example, cued the word "steel" with a probability of .00; the auditory cue "rhymes with eel" also cued the target word with a probability of .00. The combined cue, "a building material that rhymes with eel", cued the target with a probability of 1.00 *without prior learning*, even though the expected probability of the cue being effective was 0.00 (using the formula $p_a + p_b – [p_a \times p_b]$) (Rubin, 2006, p. 285). Based on the characteristics of the ballad form, and a certain amount of experimental evidence as cited above, Rubin found that in the ballad form at least, recall is serial; what is sung cues what is to come. Ongoing cues are based on poetic devices, including rhyme, alliteration, and assonance; meaning, visual imagery, and spatial imagery, which also function in a local, serial fashion; and rhythm, the only ongoing cue of a global associative nature (Rubin, 2006, p. 287). Rhythm functions through repetition of a near-identical rhythmic pattern repeated throughout the verses. Multiple cues combine to constrain the number of possible word choices in any particular context. As cited above, a particular word can be retrieved from weak rhyme and meaning cues when the cues are used in combination.

Up to this point, I have presented two rather distinct views of song memory. The first is derived from studies of expert piano performance (Chaffin & Imreh, 2002) and self-observation of a singer/psychologist in preparation for a performance of a short work by Stravinsky (Ginsborg & Chaffin 2007). As such, this view may have more relevance to the study of expert memory than as a view of song memory in the general population. This view presents song memory as content-addressable through declarative performance cues. Rubin and Wallace present a different view based on observations and analysis of ballads and the oral tradition. They make a compelling case for song memory that relies on the surface features of the music and text to cue an ongoing associative retrieval in performance. It is not content-addressable; once interrupted the performance must start over from a section break or from the beginning. As such, the focus has shifted from the role of semantic schemas for memory retrieval to an understanding of the role of multiple schemas in multiple separate episodic systems, each contributing their unique part to the memory of the learning event. More particularly, the key to memory for song may be in serial recall through a system of multiple constraints, whereby surface features may act as a primary mechanism for long-term retention and retrieval. Following Rubin (2006), this process relies on multiple systems within episodic memory for its effect. Because spacing learning events can maximize the contributions of episodic memory to the effective coding and retrieval of a wide range of materials, it could also have an effect on the performance of song.

2.6 The spacing effect.

The distributed practice or spacing effect refers to the retrieval advantage conferred when information is presented in study sessions distributed in time. This is usually compared to the same information studied for the same amount of time in a massed format. Figure 2 illustrates a design for a typical distributed practice study. A massed session is where all of the learning trials are conducted without a break, or where there is a short break on the order of seconds. In a spaced session the break between the trials, or *inter-study interval* (ISI) may be minutes, hours or days later. The *retention interval* (RI) is the time between the second study event and the final test. Distributed practice may also refer to the distribution of spaced learning episodes with no massed condition, the format used in this study. In the usual spacing study, there are two study events and a final test. The first session presents the material for learning, often to a criterion of 95% correct recall at the end of the session. The second session usually consists of an initial test and a fixed number of relearning trials. If the second session trials are not fixed, the study is vulnerable to confounding between learning and memory. Without fixed second session trials, some participants could systematically receive extra learning for the longer ISI conditions over the shorter ISIs because of the increased forgetting at longer intervals. Under these circumstances the difference in final results may be at least partially due to variable learning and not to the spacing interval alone. The retention interval (RI) is a typically fixed interval, which allows the ISI to be examined as a single independent variable. For convenience, the terms spacing effect and distributed practice effect are often used interchangeably.

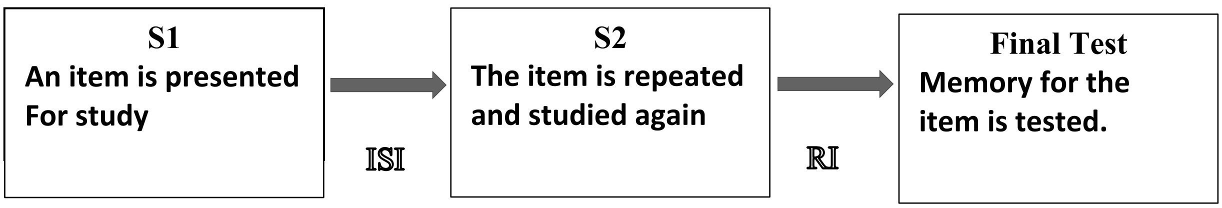

Figure 2. A basic distributed practice design (based on Wiseheart et al. 2019)
During the first study event, new material is introduced and learned. During the second study event, the material is reviewed. During the test event, individuals are tested on their memory for the material. The time interval that passes between the first and second study events is referred to as the interstudy interval. It is either (a) short/massed or (b) long/spaced. The time interval that passes between the second study event and the test event is referred to as the retention interval. It is either (c) short or (d) long. Note: Some researchers refer to interstudy interval as gap or lag. (Wiseheart et al., 2019)

One of the interesting aspects of the spacing phenomenon is the interaction between retention interval and interstudy interval. The time between study events necessary for optimal memory retention changes according to a U-shaped function, (the Glenberg surface) depending on the time elapsed between final training and testing (Glenberg, 1979). In a study dealing only with verbal materials presented in paired-associate lists, Glenberg (1979) found that the optimal interstudy interval for a given RI increased until a certain critical point. After that, performance on the final memory test decreased. The curve of this was characteristically U-shaped and non-monotonic; that is, the optimal ISI increased to a certain point and then decreased. Later work clarified the optimal ISI for simple verbal materials for a wide range of time scales (Pashler, Cepeda, Lindsay, Vul & Mozer, 2009; Küpper-Tetzel & Erdfelder, 2012).

Glenberg's explanation of encoding variability (1979) outlines the two different principles that underlie memory performance:

> First, a repetition is potentially effective to the degree that the second presentation allows for the storage of information distinct from that stored at the first presentation. Second, the realization of this potential is controlled by the conditions at the time of the memory test, the retrieval environment. (Glenberg, 1979, p. 96)

In a spacing context, successful retrieval at the second learning episode creates a stronger memory trace for the item, a process termed *study-phase retrieval* (Delaney, Verkoeijen, & Spirgel, 2010; Cepeda et al., 2006). A corollary to this theory is that if too much time elapses between the study events, the item may be forgotten. In this case, there is no strengthening of the initial memory trace; the learning material is instead encoded as a new event, and there will be no study-phase retrieval. The theory suggests that the most effective interstudy interval will be determined by the nature of the material to be learned and the particular characteristics of the learner, but it cannot account for the specific factors that determine success (Wiseheart et al., 2019). It also cannot explain why the space between the learning events must increase as the retention interval increases to insure optimal retention. Because of this, a second theory is needed to fully account for experimental findings.

The theory of *contextual variability* accounts for the effect of ISI on final test performance. A change in time between study sessions alters the number of largely unconscious context-dependent cues available at the time of testing. When an item is studied, it is encoded along with the cues available from the learning environment. These cues include all the episodic information that identify the particular learning event, including external elements such as the physical environment, time of day, presence of the experimenter and the internal characteristics of the learner, including emotional state, prior schematic understanding, and individual learning strategies (Glenberg 1979). When contextual cues at the second learning episode are identical to the cues encountered at the first episode, the memory for material encoded alongside those cues is also strengthened. Newly encountered cues also form new associations with the material to be remembered. The increase in contextual elements absorbed in the learning context increases the likelihood of an overlap with the contextual cues available at the time of the final memory test. Spaced repetitions allow the encoding of a greater number of different contextual elements than the massed sessions, where the context throughout the learning sessions is more similar. This means that there will be fewer unique contextual cues associated with the material from the massed sessions, and less chance of encountering those same cues at testing. Eventually the change in time between study sessions reaches a certain optimal point for any given retention interval. Beyond that, the increase in contextual elements is overtaken by the drift in context at time of testing away from the context of the learning sessions. The context at testing then shares fewer contextual cues with the learning sessions, and the material is less likely to be retrieved.

These complementary theories fit well with the new theory of disuse cited above (Bjork & Bjork, 1992). In Bjork and Bjork's terms, the spacing effect depends on the interaction between the storage strength of an item and the retrieval strength of the cues. Storage strength is a measure of how well-learned an item is. Retrieval strength describes the accessibility of an item in memory in performance, when given a specific set of retrieval cues. When items are harder to retrieve (due to increased length of time between learning sessions) the storage strength increases, provided retrieval is successful (study-phase retrieval). When more contextual elements describing the event are absorbed in the learning phase, retrieval strength of the cue is increased, provided the cues are relevant to the contextual environment at the time of testing (contextual variability).

Encoding variability was only one part of the components-level theory advanced by Glenberg to explain the spacing effect in verbal memory. There are three types of cue considered to be significant in recall, contextual, structural-associative and descriptive (Glenberg, 1979, p. 96). Of these, contextual cues function largely unconsciously, based on the particular context of the learning environment, which is automatically encoded. Structural cues are based on the structural relationships discovered or imposed on items to be remembered during the learning phase. They include the schematic structures imposed on words during study (mnemonic strategies, or the repetitive rhythmic patterns of poetry, for example) that can enable retrieval. They are more specific than the contextual cues which are typically general in nature. Finally, descriptive cues are meanings attached to the material and are the most specific in nature. They represent the unique features of the material to be learned and are used to differentiate it from all other material. They include, for example, all the lexical features of the spelling of a word, as well as its particular sound and meaning. Cues that share these components can specifically target the learned material for retrieval. Generally speaking, components of a more specific nature take precedence over components of a less specific nature when they are activated by a given cue.

The spacing effect has been demonstrated repeatedly over the last 125 years for syllable lists, sentences and passages of text across the life-span (Cepeda, Pashler, Vul, Wixted & Rohrer, 2006). Distributed learning has been much less studied in difficult conceptual tasks and tasks involving complex motor co-ordination (Wiseheart et al., 2019). Spacing has a large positive effect on verbal learning (d = .85; Cepeda et al. 2006; Moss, 1995)[4]. Motor skill learning shows a medium positive effect (d = .5; Donovan and Radosevich, 1999; Lee & Genovese, 1988), as

[4] Cohen's *d* is a measure of the magnitude of a phenomenon.

does intellectual skill learning (d = .5) (effect size estimates from Wiseheart et al., 2019, based on Foot, 2016, Kapler, Weston, & Wiseheart, 2015, and Vlach & Sandhofer, 2012). Social and emotional skills show a small positive spacing effect (d = 0.2) (effect size estimate from Wiseheart et al., 2019, based on Korben, 1976 and Rowe & Craske, 1998).

Music learning and the spacing effect. Playing an instrument requires multiple finely-coordinated physical actions and necessitates multiple practice sessions to acquire the particular skill (Simmons, 2012). Learning music requires the simultaneous coordination of visual input (reading notation), motor output, auditory imagination, aural monitoring and modification of motor behaviour through auditory feedback (Simmons, 2012, p. 3). Using a short left-hand piano figure learned and tested from score, Simmons and Duke (2006) found significantly fewer performance errors for an interval separated by sleep at RIs of 12 or 24 hours than for an RI of 12 hours without sleep. This was a single session study with a varied retention interval, not a distributed learning study. Using a similar left-hand piano figure read from score, Simmons (2012) found fewer performance errors at an RI of 24 hours for a lag of 24 hours compared with a lag of 6 hours. However, the study did not train participants to a uniform criterion of errors in the first session, making it impossible to separate the effects of differential learning from the effect of the inter-study intervals. In addition, results were an average over sessions, without a retention interval. Under these circumstances, benefits from spacing could not be determined. Rubin-Rabson (1940) evaluated learning of short piano pieces among experienced pianists. The method used allowed a variable number of trials at the second learning session, so the effect of longer lags was confounded by a greater number of learning trials. Cash (2009) studied the effect of a 5-minute break on learning a keypress sequence or a sequence of 13 notes. Results showed improved performance for an early five-minute gap over a later five-minute gap when tested 12 hours later after sleep. While not a memory study, it did demonstrate the positive effect of a gap on mechanical accuracy. A study by Wiseheart, D'Souza and Chae (2017) using five different lags between 0 and 15 minutes, found no spacing effects at a retention interval of 5 minutes, either for key press with visual directions or memorized song fragments. The data indicated no forgetting had taken place before the second learning sessions, so study-phase retrieval could not occur. This leaves open the question of what effect an interval of sleep between learning events would have on a subsequent measure of memory performance. Studies by Stambaugh (2011) and Stambaugh and Demorest (2010) examined short phrases played on clarinet or saxophone for accuracy and musicality in a massed or interleaved condition. They were not memory studies and did not use a lag between study events, limiting their applicability to this research.

There is currently no literature on the effects of distributed learning on song memory. Because of its reliance on implicit context-based associative phenomena, spacing as a learning technique could have a powerful role to play in the long-term retention of song. The optimal distribution of song learning among a normal population is relevant as a comparative baseline for musical mnemonic training, a therapeutic approach to memory (Murakami, 2017). A recent study of singing as an aid to foreign language learning in a population of school age children (Good, Russo & Sullivan, 2014) used a learning schedule of six sessions distributed over two weeks. This was a choice based on the hands-on experience of teaching English as a second language. There is currently no empirical research to support their decision. Singers preparing for performance currently have no idea what practice schedule they should use to better guarantee fluent memory retrieval during performance. This research offers an empirical and theoretical justification for a commonly-encountered but as yet unstudied aspect of sung performance.

Before beginning the present study, I conducted a small pilot study (n = 4) using similar materials and design. The four participants were advanced singers drawn from the second-year class of an undergraduate program at a local music conservatory. The pilot study assessed syllable and pitch errors in tests cued by first notes, then first notes and syllables of a two-verse song adapted from traditional materials, after spacing intervals of five minutes and one week, tested at a retention interval of six weeks. Melody was retained accurately by both groups; but the spaced group showed mean correct syllable scores of 94.5% for the spaced and 29.7% for the massed condition cued by first notes, and 94.1% for the spaced and 58.3% for the massed condition cued by first notes and words. Results indicated a large difference in mean correct syllable scores between the two groups, attributable to the spacing effect. The pilot testing gave me reasonable confidence that a two-verse song based on traditional materials could be learned to criterion in a single session. Further materials testing determined that the original choice of lyrics, the published lyrics to the song "Come all ye old comrades" (Creighton & Senior, 1950) were unsuitable for university undergraduates due to subject matter and archaisms in the text. The lyrics were rewritten accordingly to make the song more interesting to the target population.

2.7 Research questions and hypotheses

My research questions in this dissertation are whether there is an optimum distribution of learning sessions for song (analogous to that for spoken or written syllables), that could help singers' memory for sung verbal materials; and whether the effect of that optimal distribution will be maintained over an increasing distance between learning events. While distributed practice enhances free recall and comprehension for substantial sections of text (Rawson & Kintsch, 2005), it is not directly comparable to memory for text set to music, where continuous, verbatim recall according to an imposed rhythmic and melodic pattern is required. Furthermore, in sung verbatim recall, the procedural aspects of singing technique and the demands of music processing are engaged; the rhythmic and prosodic constraints of the poem and the rhythmic and pitch constraints of the tune function as a framework for the song materials (Wallace & Rubin, 1991). This framework is presented complete at the first learning episode. When prosodic aspects of the poetic form including stress patterns, rhyme, alliteration, and verse structure are understood and the musical pitch and rhythm learned, the framework is in place. The words are associated with the rhythmic, prosodic and melodic pattern through repetition. Following Glenberg (1979), remembering a song requires retrieval of the episodic traces representing exposure to the song in the learning phase of the experiment. Access to these traces is provided by the cue at testing. The cue allows for activation of components in the episodic trace identical to those in the cue. In the case of song, structural components created by the poetic and musical framework internalised during the learning phase may be activated when the first notes or words are heard. Access to the memory is predominantly controlled by the most specific components in the trace (Glenberg, 1979). Since structural components are more specific than contextual components, the access to song memory will be controlled by the structural components implied by the initial cue, and the ongoing associative cues generated by the performance as it unfolds. The effect of contextual variability is diminished (but not eliminated) but the associative pattern generated by the material itself takes precedence. With short spacing intervals, study-phase retrieval relies on the easy retrieval of information from the first occurrence. As the spacing interval increases, study-phase retrieval relies on a ‘reconstructive process’ (Glenberg, 1979, p. 110) that is closer to the demands of the final memory test. In song, this reconstruction is aided by the structural information offered by the cue. There is little gain to the storage strength of the song from wider spacing intervals, because retrieval of highly structured material is not more difficult at intermediate intervals of differing lengths, provided sufficient structural information can be recovered from the cue. For this reason, there should be a spacing effect for sung poetic materials, but a minimal difference between two spaced learning episodes, when compared at the same retention interval.

The following a priori hypotheses were made:

Hypothesis 1: Memory for an unaccompanied song based on traditional materials will show a spacing effect in note and syllable accuracy when an interval of ten minutes between two learning sessions is compared to spacing intervals of two days and one week, at a retention interval of three weeks. A song based on traditional materials introduces a wealth of associative elements related to on-going constraints from the prosodic structure, imagery, rhyme, and metre of the text; the metrical and pitch characteristics of the musical setting; and the narrative and affective nature of sung material (Wallace & Rubin, 1991). The storage strength of the words is related to associations formed between the words and these individual cuing features of the song. Study-phase retrieval should allow for an increase in storage strength of sung text in a spaced over a massed condition. The increase in contextual cues offered by the changed learning environment at the spaced sessions will also contribute to improved retrieval in the spaced conditions.

Hypothesis 2: Memory for an unaccompanied song based on traditional materials will not show evidence of difference in note and word accuracy when study intervals of two days and one week are compared at a retention interval of three weeks. The structural cues generated internally by the song materials are enough to overtake any additional boost to learning between the two spaced sessions offered by increased contextual variability, once the material is stored with sufficient associative strength. Any tendency toward decrease because of too much space between the spaced intervals in relation to the retention interval would also be equalized. For this reason, there should be minimal difference in error scores for the words of a traditional song between two spaced learning episodes, when compared at the same retention interval.

Hypothesis 3: There will be evidence of a greater number of errors in note and syllable recall at the initial second session tests, when spacing intervals of two days and one week are compared to a spacing interval of ten minutes. At the ten-minute interval, material to be remembered will show minimal forgetting, thanks to the strength of the initial memory trace at a short interval, and the strong match between contextual cues at learning and at testing. The two spaced intervals will show substantial forgetting due to weakening of the initial memory trace and the mismatch between contextual cues encountered during learning and those encountered at retrieval.

Hypothesis 4: There will be no evidence of a greater number of errors for notes and syllables at the initial second session tests, when the two intervals spaced at two days and one week are compared. Memory for the material at the spaced sessions will be largely determined by the cue strength: the degree to which the constraining patterns associated with the song can be recovered from the cue will determine what can be remembered. The difference in context at the second session which would normally govern the decline in retrieval of well-learned material[5] at increasingly spaced intervals will be superseded by those constraints in the material that can be recovered from the cue.

---

[5] In this case, material learned to criterion.

3. Methodology

3.1 Participants

One hundred and eleven York University undergraduates were recruited. Participants in the fall term (*n* = 91) were drawn from a second-year music skills class for music majors and received course credit for participation in the study. A further group in the spring term (*n* = 20) were recruited by poster from the general university community and given a coffee card as incentive. Participants were randomly assigned to one of three groups, a group of 36 with a 10-minute ISI, a group of 39 with a two-day ISI and a group of 36 with a one-week ISI.[6] The sample size was chosen to provide 85% power, based on an estimated large effect size from previous meta-analyses of the spacing effect in verbal learning (d = 0.85) (Cepeda et al., 2006). Participants were excluded from analyses if they were above the cut-off age (34 or older); had learning disabilities; or for missing or overhearing sessions. Of the remaining participants, 87 were able to reach a criterion of 95% correct syllable retrieval in the first session and completed the study. In total the sample consisted of 45 females and 42 males distributed as follows: the massed (M1) group was composed of 20 females and 10 males, $M_{\text{age}} = 20.97$, $SD_{\text{age}} = 2.49$; the S1 group 11 females and 18 males, $M_{\text{age}} = 20.97$, $SD_{\text{age}} = 3.448$; the S2 group 14 females and 14 males; $M_{\text{age}} = 20.93$, $SD_{\text{age}} = 3.25$. Bayesian ANOVA indicated evidence of no difference between conditions for age ($BF_{10\ \text{overall}} = 0.102$) (see page 54 below for a brief explanation of Bayesian statistics).

---

[6] A study of complex materials such as this cannot present a truly massed condition, because the initial learning session involves many breaks between each presentation of the stimulus materials. For convenience the ten-minute ISI was chosen to represent the massed condition and labelled “M”. The two-day and one-week intervals were the spaced conditions and were labelled “S1” and “S2” accordingly.

Table 1

Native language by experimental group

Language	Massed	Spaced 1	Spaced 2
Bengali	0	1	0
Chinese	3	2	2
Creole	0	0	1
English	26	20	20
Greek	0	0	1
Hindi	1	1	0
Korean	0	1	1
Portuguese	0	1	1
Spanish	0	2	2
Tamil	0	1	0
Total	30	29	28

Twenty-six of the participants were native English speakers in group M1, 20 in group S1 and 20 in group S2, Bayesian analysis indicated no evidence of difference between groups ($BF_{10\ overall}$ = 2.054) (Table 1). Average number of years of spoken English were *M* = 19.80, *SD* = 4.13 in group M, *M* = 19.10, *SD* = 5.24 in group S1 and *M* = 19.32, *SD* = 4.28 in group S2, with Bayesian analysis indicating evidence of no difference. (BF_{10} = 0.118). There was evidence of no difference between groups for bilingualism (BF_{10} = 0.104) (rated from 1 to 10).

Table 2

Voice type by experimental group: Soprano, Mezzo-Soprano, Alto, Tenor, Baritone and Bass

Condition	S	M	A	T	BAR	B	Total
M	7	6	7	7	2	1	30
S1	3	4	4	5	9	4	29
S2	3	1	10	6	3	5	28
Total	13	11	21	18	14	10	87
BF_{10} independent multinomial							1.976
N							87

Bayesian analysis indicated no evidence of difference between groups ($BF_{10\ \mathrm{overall}}$ = 1.976) for voice type (Table 2).

Table 3

Participants who self-identified as being able to sing

Frequencies for singer

Condition	Singer	Frequency	Percent	Valid Percent	Cumulative Percent
M	n	8	26.7	26.7	26.7
	y	22	73.3	73.3	100.0
	Missing	0	0.0		
	Total	30	100.0		
S1	n	9	31.0	31.0	31.0
	y	20	69.0	69.0	100.0
	Missing	0	0.0		
	Total	29	100.0		
S2	n	6	21.4	21.4	21.4
	y	22	78.6	78.6	100.0
	Missing	0	0.0		
	Total	28	100.0		

Years of voice lessons. There were some differences between groups for years of vocal training. The massed group had M = 2.44, SD = 3.78, group S1 had M = 0.50, SD = 1.16, and group S2, M = 1.40, SD = 1.92. Bayesian analysis showed there was evidence of difference between conditions M and S1 (BF_{MvsS1} = 4.506). However, there was there was evidence of no difference between conditions for number of performances of song from memory ($BF_{10\ \mathrm{overall}}$ = 0.105); strong evidence of no difference between conditions for the number of participants who self-identified as being able to sing ($BF_{10\ \mathrm{overall}}$ = 0.095) (Table 3); and evidence of no difference between conditions for number of hours singing per week ($BF_{10\ \mathrm{overall}}$ = 0.157). Twenty-two of group M1, 20 of group S1 and 22 of group S2 considered themselves to be someone who could sing.

There was no evidence of difference between conditions for hours of instrument playing per week (BF_{10} = 0.873) or for years of instrumental training (BF_{10} = 0.352). Twenty-five played at least one musical instrument in group M1, 29 in group S1 and 25 in group S2. There was strong evidence of no difference between conditions for method of music learning, whether by ear, by sheet music, or both (BF_{10} = 0.014). There was evidence of no difference between conditions for sheet music reading proficiency (BF_{10} = 0.234) (rated on a scale of 0-10) or number of times performing a song from memory (BF_{10} = 0.105). There was strong evidence of no difference between conditions for number of voice or theatre majors compared to others (BF_{10} = 0.074). Finally, there was evidence of no difference between conditions for how anxiety affected performance (BF_{10} = 0.111) (rated on a scale of 0-10) and evidence of no difference between conditions for the PROMS scores of music perception individually or overall (BF_{10} = 0.168). The experiment was approved by the Research Ethics Board of York University and all participants gave written consent prior to beginning.

3.2 Apparatus

Three laptop computers were used. The first, a 2.6 GHz MacBook Pro (2013), recorded the stimulus and the participant files in Logic Pro; the second, a Dell laptop, presented the stimulus materials in a PowerPoint format to the participant, using an external VGA monitor. The third Dell laptop was used at the final session for the on-line PROMS test of musical perception abilities. A Steinberg UR 242 microphone positioned 18 inches to the right of the participant and 24 inches to the left of the experimenter was used to record both the participants and any remarks from the researcher. The microphone output was connected to a three channel PreSonus MK II mixer, which was then connected to the Mac laptop for recording.

3.3 Materials

A song was newly composed by the author based on "Come all ye old comrades", song 59 of the Traditional Songs from Nova Scotia (Creighton & Senior, 1950) (Figure 3). Efforts were made to respect and enhance the melodic simplicity, rhythmic regularity, consistent rhyming structure and concrete textual imagery characteristic of songs in the oral tradition (Wallace 1994, p.1473). The original melody was in the Dorian mode, centering around the second scale degree, and had a wide range, including leaps of a seventh. The recomposed tune reflects a very simple underlying tonic-dominant harmony (I-V-I-IV-V-I) and a range limited to a minor seventh. The melody is largely conjunct for ease of vocal emission but allowed several leaps from the dominant to the tonic tone for melodic variety. The ¾ rhythm was kept, as was the pattern of four four-bar phrases with an upbeat.

Melodies using pitches that outline tonic triads at phrase endings are remembered more accurately than comparable melodies without, but only when those triads are employed within regular accent patterns (Patel, 2010, p. 203; Boltz, 1991); and repetition of similar elements aids chunking in memory (Bellezza, 1996). In an attempt to make the tune easy to learn, I incorporated a motif of four repeated notes in a rising pattern at the beginning of the first three phrases, with a rounding off phrase down to the lower dominant tone before returning to the tonic note at the final cadence. The overall structure is that of three rising melodic arches and a balancing descending phrase returning to the opening tonic pitch (Figure 4). Materials testing indicated that words based closely on the original text were not well understood or remembered by undergraduates, largely due to archaisms and the unfamiliar context of the original, a maritime song dating back to the 18th century (Creighton & Senior, 1950). In an effort to create materials more in keeping with the interests of the target population, I composed new words that respected the rhythmic profile of the original (Table 4). Each line of the new poem contained three trisyllabic, dactylic feet (/UU) with an upbeat, ending with a stress (/), for a total of four stresses in each line. Upbeats occasionally had more than one syllable. The newly composed words had the following syllable count: Verse 1: 11/12/12/12, Total 47. Verse 2: 11/12/12/12, Total 47. Total syllables: 94.

Figure 3. Original ballad tune.

Figure 4. Recomposed tune.

Table 4.

Original and recomposed versions of song text

Come all ye old comrades, come now let us join, Come and lift your sweet voices in chorus with mine For we'll laugh and be jolly while sorrow refrain For we may and may never all meet here again. Fare ye well, I had a sweetheart which I dearly loved well, Without, or with beauty, there is none to excel; She would laugh at my folly as she'd sit on my knee, There were few in this wide world more happy than we.	It's time for vacation, I studied enough, I am sick of my courses the profs are too tough; I don't have any reason to get out of bed, And the thought of my homework just fills me with dread. My data is done and my phone has no juice, No more internet dating I can't see the use; All the stress is a pain and I need some release, But my head is exploding so leave me in peace.

The song scores were prepared using Noteflight, an online music transcription software. Scores were prepared in the treble clef in the keys of F for soprano, Eb for mezzo and D for alto voice, and in the tenor treble clef in the key of F for tenor, and in the bass clef in D for baritones and C for basses (Figure 5). The songs were recorded on the piano in the appropriate key and octave for the different voice types by a pianist in a professional performance program at a local conservatory, using a click track set to quarter note = 138, a tempo chosen to sound natural with the words. The songs were then recorded by three singers (undergraduates in a professional vocal program) using a click track set to quarter note = 138, with the Steinberg microphone and preamplifier in LogicPro software using the MacBook computer. The soprano recorded the material twice, once in F and once in Eb. The Eb version was transposed into D to produce the alto materials using Melodyne, a professional note-editing program. The tenor version (in F) was recorded separately, as was the baritone (in D). The baritone performance was transposed one tone down into C to generate bass materials.

Six different PowerPoint presentations in the different stimulus keys were then prepared using the stimulus recordings. A script was written for each of the three sessions, with the stimulus recordings embedded in the presentation (see Appendix 1 for a sample script). The researcher then recorded instructions and prepared instructional slides to go with the recordings. All participants in the study were given the mini-PROMS (profile of music perception skills) test (Zentner & Strauss, 2017), a 15-minute version of the original PROMS battery of tests. The original PROMS battery included a reliable, well-validated set of tests comprising melody, tuning, tempo, rhythm, embedded rhythms, accent, pitch, timbre, and loudness, for a total of 9 different subtests (Law & Zentner, 2010). The mini-PROMS contains only four subtests:

melody, tuning, tempo, and accent. The mini- PROMS showed a test-retest reliability of $r = .83$ ($P < .001$) indicating that the test results were highly replicable; criterion validity, determined through correlation with years of musical training, musical qualification, and musicianship status, was $r = 0.61$ ($P < .001$), indicating a moderately high correlation of test results with demographic measures of musical expertise. The overall mini-PROMS score tests showed a moderately strong correlation of $r = 0.52$ and $r = 0.48$ with convergent validity scores determined for the 9-test PROMS-S battery[7]. While the mini-PROMS has not itself been examined for discriminant validity, discriminant validity for the PROMS-S and a test of fluid intelligence, Raven's APM (Advanced Progressive Matrices) was r = .27, showing a modest correlation of the PROMS-S and intelligence consistent with previously reported data. Taken together these findings suggest that the both PROMS-S and the mini-PROMS drawn from it do indeed measure musical ability. Results indicated evidence of no difference between groups for musical perceptual ability ($BF_{10\ overall} = 0.168$).

[7] Using a musical competence scale, separately administered ($\beta = .56$, $p < .001$),

Figure 5. Soprano version of stimulus materials.

3.4 Procedure

The participants were first exposed to the two-verse song melody played on the piano in a key appropriate to their voice type. After the song melody was presented and imitated in line-by-line, phrase-by-phrase and complete form, participants attempted to sing the song from score.

Where the performance was judged by the researcher to be error-free they continued on to text presentation and text with melody learning. Where mistakes were made, participants were returned to more presentations of the melody and another melody test from score. Participants who had returned to the melody learning trials were given a maximum of three more exposures to the complete song melody and then tested again. All participants were then exposed to the learning trials for the song with words. (Figure 6)

Training to Criterion

Melody practice session 1	**Melody test session 1**	**Song practice session 1**	**Criterion test session 1**

Inter-Study Interval

Pre-practice test 1 session 2	**Pre-practice test 2 session 2**	**Song practice session 2**	**Post-practice test 1 session 2**	**Post-practice test 2 session 2**

Retention Interval

Final test 1 session 3	**Final test 2 session 3**	**Relearning to criterion session 3**

Figure 6. Study design

After introductory slides, the song recording was played without the score. Presentation of the song with words proceeded similarly to the presentation of the tune, this time using the vocal recordings and the score with words. The song with words was presented and imitated in line-by-line, phrase-by-phrase and complete form, until participants were ready to try the song from memory. Those who indicated they were not ready were then coached by the researcher so that they might reach criterion within the allotted maximum session time of 45 minutes. At this point participants were returned to the three presentations of the compete song with score. This cycle of deciding whether or not to do the memory test, coaching and being returned to learning trials continued until the participant was ready for the memory testing. The final test of the first session was cued by the tone indicating the initial pitch and the first two notes and words of the song from the demonstration recordings. Testing and exposures to the stimulus recording continued until criterion was reached, or the forty-five minutes allotted for the session had elapsed. Eighty-seven of the 100 qualified participants reached criterion. At the end of first session testing, a demographic questionnaire was completed. Massed (10-minute ISI) participants were engaged in conversation for the remainder of the 10-minute gap between sessions to prevent mental review of the materials. Participants in the spaced conditions were thanked for their participation and reminded of the second appointment.

The second session procedure was uniform across all conditions. Participants were given two initial memory tests using the identical format to the memory tests from the first session. Tests were cued by the two first notes played on the piano and then by the two first notes and words of the song. Where participants paused for 10 seconds or more, or otherwise indicated hesitation, they were queried by the researcher for "anything else?". Where participants began singing notes instead of notes and words, they were stopped and asked to sing notes and words, as directed by the slides; all notes and words sung or spoken, however, were included in the analysis, as long as they were in the rhythm of the poetic text and not a paraphrase. After the memory tests, participants were exposed three times to the stimulus materials, and instructed by slide to sing along. They were then given the final memory tests for the second session, thanked and reminded of their final appointment in three weeks.

At the third testing session, tests with melodic and melody and word cues were given. Although sung performance of text was requested, credit for any correct words spoken in the rhythm of the poem was also granted. Any notes sung without words were also included. Participants who did not reach criterion in the final test were given further training with the recording and score and tested after each training until they reached criterion for the words. All participants were then tested with the shorter PROMS battery to establish levels of music perception skills.

Although testing was conducted by the lead researcher, who was not blind to condition, efforts were made to ensure freedom from bias and to establish equivalence between participant groups other than for the experimental manipulation. Of the three sessions in the lab, sessions 2 and 3 followed a strict protocol determined by the slide presentation. The initial session necessitated individualized coaching in order to reach the criterion learning goal. All the individual first sessions were compared post hoc to determine relative equivalence in coaching styles between groups. Forty-two participants out of the 87 who reached criterion were prompted to a verbal recitation of the song text. There was, however, evidence of no difference between groups for time spent in verbal recitation ($BF_{10\ overall} = 0.153$)[8]. There was some variation in vibrato, portamento, quality, pitch and rhythm in the stimulus recordings depending on the personal characteristics of the singers and the degree of post-production editing by the researcher. A Bayesian frequency analysis indicated no evidence of difference in distribution of stimulus materials across experimental conditions ($BF_{10\ overall} = 1.976$). Examination of the final performance from the first learning session (test 1.3) showed evidence of no difference between groups for cents off-pitch quarter notes ($BF_{10\ overall} = 0.174$); *SD* of cents off-pitch quarter notes

[8] See note 9 for an explanation of Bayesian frequency analysis.

($BF_{10\ overall}$ = 0.220); absolute value (*ABS*) of cents off-pitch quarter notes ($BF_{10\ overall}$ = 0.198); number of off-pitch quarters ($BF_{10\ overall}$ = 0.216); or proportion of quarter notes off-pitch ($BF_{10\ overall}$ = 0.215). Post-hoc examination of learning outcomes also showed evidence of no difference overall for mean quarter note length at the final first session test ($BF_{10\ overall}$ = 0.105) and no evidence of difference for the final second session test (Test 2.4) ($BF_{10\ overall}$ = 0.819). From this it can be assumed that all participants across conditions learned the tune and the words to the same standard of correctness.

3.5 Data Collection

Session files in .wav format for each participant were uploaded to Dropbox. These were then converted into a blinded format by the Dissertation supervisor so that coding was done anonymously and reposted to Dropbox. From there, the session were downloaded to Alchemy, an open source DAW (digital audio workstation), where seven tests were separated from the original session files for analysis: the criterion melody test and final criterion test from the first session; the initial memory tests and final memory test from the second session; and the memory tests from the final session, cued by initial notes, and then notes and words. Separated test files were labelled and re-uploaded to Dropbox.

Test files were then downloaded by the author into Melodyne, a note-based audio processing software. Melodyne operates by detecting regions of pitch stability ('note objects') and the sudden amplitude changes characterising note onsets and offsets ('event objects') (Neubäcker, 2011). These regions are then graphically represented by note blobs in the display, with each blob having a stable series of harmonics corresponding to the detected average note frequency. The frequency of each time interval examined is determined through a mathematical function, the fast Fourier transform (FFT). Because note objects have a stable periodicity, they can be separated from noise, which is non-periodic in frequency.

The software used pitch and amplitude information to separate the test files into separate notes according to its proprietary algorithm. These assigned notes were then checked by the lead researcher by ear and readjusted where necessary to correspond to what was actually heard. Where participants anticipated a later phrase, then returned to the normal sequence of the song, the initial anticipation was not coded, but treated as a hesitation. Where participants sang a note in sequence with the wrong text and then corrected the text, the result was scored from the correct syllable, with the initial attempt treated as a hesitation. Notes with a long portamento were split so that the first half of the note with the slide did not affect the average pitch of the note. The average pitch was then taken from the arrival point. Note times, however, were taken from the note onset. Occasionally pitches were stable in the first part, then unstable in the second. This was judged by a wide vibrato on the note, or a simultaneous decrescendo and pitch change, indicating a loss of vocal control. In these cases, the note was split, and both pitch and onset information taken from the first half of the note. Ten percent of the note assigned files were checked by a second rater (Table 5). Once the notes were assigned in the Melodyne files to the satisfaction of the lead researcher, the algorithmically generated values for pitch (in note names and cent deviations) and note length, breaths and hesitations (in hundredths of a second) were then transcribed by one of three different coders and entered into spreadsheets on Google Drive.

Data collected at the sessions allowed examination of syllable memory, pitch memory, and length of notes, breaths and hesitations. Syllable memory was reported in number of correct syllables, number of syllable omissions, number of additions, and number of incorrect syllables recalled. The number of relearning trials to reach syllable criterion in the third session were also tracked. Pitch data was reported in number of notes omitted; number of pitched quarters, dotted quarters and halves, and eighth notes; mean cents off-pitch and standard deviation (*SD*), absolute value (*ABS*), skewness and kurtosis of means cents off-pitch for quarter notes; and number of quarter notes off-pitch. Note-lengths, breaths, and hesitations were reported in hundredths of a second. Results were analyzed with Bayesian ANOVA and individual Bayes factors calculated to confirm main effects and interactions.[9]

Table 5

Inter-rater reliability correlations (Pearson's r)

Measure	Test 1.1	Test 1.3	Test 2.1	Test 2.2	Test 2.4	Test 3.1	Test 3.2
Word omissions	NA	1.000	0.999	0.999	0.986	0.999	1.000
Note omissions	NA	1.000	0.999	1.000	0.986	0.999	1.000
No. cents off quarters	1.000	0.999	0.966	0.999	1.000	.899	.888
Quarter length	0.959	0.983	0.783	0.964	0.984	0.951	0.972
Word errors	NA	0.958	0.991	0.922	0.969	0.940	0.924

[9] "Null hypothesis significance testing (NHST; e.g., ANOVA) tests only whether to accept or reject a null hypothesis (H_0); it implies nothing about an alternative hypothesis (H_1). On the contrary, Bayesian statistics test the probability of the observed data under H_1 relative to H_0, providing a richer interpretation of the data. For a simple interpretation, a BF value of 1 means the data are equally probable under H_0 and H_1. A BF value less than 1 means the data are more probable under H_0 relative to H_1 (0.33-1 = inconclusive evidence for H_0; 0.1-0.33 = moderate evidence for H_0; <0.1 = strong evidence for H_0). A BF value greater than 1 means the data are more probable under H_1 relative to the H_0 (1-3 = inconclusive evidence for H_1; 3-10 = moderate evidence for H_1; >10 strong evidence for H_1). For the interested reader, see Jarosz and Wiley (2014)." (Footnote from Weston, 2018, p. 48)

Results

Results were measured at seven separate tests. First session tests included tests 1.1 (the final melody trial) and the final session test, 1.3 (the final criterion trial for the song, cued by notes and words. Second session tests included the initial tests, 2.1 (song cued by first notes) and 2.2 (song cued by first notes and words), and the final session test, 2.4 (song cued by first notes and words). Third session tests included the initial tests, 3.1 ((song cued by first notes) and 3.2 (song cued by first notes and words). The number of relearning trials to reach criterion after test 3.2 were also tracked.

4.1 Learning outcomes

At the end of the first learning session there was evidence of no difference between groups for syllable omissions ($BF_{10\ overall} = 0.125$); syllable errors ($BF_{10\ overall} = 0.145$); or for syllables added (no words added in any condition). Overall the song was learned to the same (correct) standard across groups (Tables 6 and 7). Whatever differences there were in pitch or timing of stimulus materials had no effect on the learning outcomes (tests 1.3 or 2.4) as measured by Bayesian analysis of any of the variables measured. There was no evidence of difference between conditions for length of first session ($BF_{10\ overall} = 0.484$) and evidence of no difference for amount of first session post-criterion learning ($BF_{10\ overall} = 0.175$).

Table 6

Learning outcome: pitch (test 1.3) Massed, 2 day, and 1 week ISIs.

Measure	Massed *M* (*SD*)	2 day *M* (*SD*)	1 week *M* (*SD*)	BF_{10} Overall	BF_{10} Massed vs. 2 day	BF_{10} Massed vs. 1 week	BF_{10} 2 day vs. 1 week
Cents off-pitch quart.	18.67 (36.81)	17.42 (31.65)	9.80 (22.77)	0.174	0.266	0.439	0.420
SD cents off-pitch quart.	31.69 (18.16)	38.53 (18.58)	35.96 (20.37)	0.220	0.618	0.358	0.297
ABS cents o.p.quart.	38.69 (26.30)	38.66 (21.65)	32.41 (12.23)	0.198	0.264	0.463	0.562

No. quar. off-pitch	19.20 (19.840)	19.07 (15.825)	14.18 (8.705)	0.145	0.324	0.323	0.270
Prop. quar. off-pitch	0.267 (0.276)	0.265 (0.220)	0.197 (0.121)	0.215	0.264	0.498	0.632

Table 7

Learning outcome: syllables (test 1.3) Massed, 2 day, and 1 week ISIs.

Measure	Massed *M* (*SD*)	2 day *M* (*SD*)	1 week *M* (*SD*)	BF_{10} Overall	BF_{10} Massed vs. 2 day	BF_{10} Massed vs. 1 week	BF_{10} 2 day vs. 1 week
Syllable omissions	0.033 (0.183)	0.069 (0.258)	0.036 (0.189)	0.125	0.310	0.266	0,305
Syllable errors	0.433 (0.653)	0.500 (0.668)	0.625 (0.997)	0.145	0.281	0.366	0.305
Syllables added	0	0	0				

4.2 Forgetting

Forgetting: syllables. In order to examine forgetting between learning sessions, tests cued by notes (test 2.1) and then notes and words (test 2.2) were used with dependent variables describing syllable memory (number of correct syllables, number of syllable omissions, number of additions, and number of incorrect syllables) note memory (number of notes omitted; number of pitched quarters, dotted quarters and halves, and eighth notes; mean cents off-pitch and SD, skewness and kurtosis of means cents off-pitch for quarter notes; and number of quarter notes off-pitch) and rhythm (note-lengths, breaths, and hesitations in hundredths of a second). Bayesian ANOVA and individual Bayes factors were calculated with learning condition as the between-subjects factor to confirm main effects and interactions.

At test 2.1 there was extremely strong evidence of difference in syllable forgetting overall, ($BF_{10\ overall}$ = 41610000), extremely strong evidence of difference between M and S1 ($BF_{10\ -MvsS1}$ = 372600), extremely strong evidence of difference between M and S2 ($BF_{10\text{-}MvsS2}$ = 2456000000), and no evidence of difference between S1 and S2 ($BF_{10\text{-}MvsS2}$ = 0.349). Mean number of syllables omitted were *M* = 1.7, *SD* = 6.64 for massed, *M* = 51.21, *SD* = 41.52 for S1 and *M* = 59.29, *SD* = 34.62 for S2 (Tables 8 and 10). There was no evidence of difference in syllable errors overall ($BF_{10\ overall}$ = 0.362) and evidence of no difference in syllable additions overall ($BF_{10\ overall}$ = 0.133). At test 2.2 there was extremely strong evidence of a difference in syllable forgetting overall ($BF_{10\ overall}$ = 4557), extremely strong evidence of difference between M and S1 ($BF_{10\ -MvsS1}$ = 499), extremely strong evidence of difference between M and S2 ($BF_{10-MvsS2}$ = 58228), and no evidence of difference between S1 and S2 ($BF_{10-MvsS2}$ = 0.353) (Tables 9 and 10). Mean number of syllables omitted were *M* = 3.4, *SD* = 11.64 for massed, *M* = 34.86, *SD* = 36.82 for S1 and *M* = 42.50, *SD* = 34.06 for S2. There was no evidence of difference in syllable errors overall ($BF_{10\ overall}$ = 0.767), and moderate evidence of no difference in syllables added overall ($BF_{10\ overall}$ = 0.164) (Tables 11 and 12) .

Forgetting: number of notes omitted. At test 2.1 there was extremely strong evidence of difference in number of notes forgotten overall, ($BF_{10\ overall}$ = 125369), extremely strong evidence of difference between M and S1 ($BF_{10-MvsS1}$ = 15987), extremely strong evidence of difference between M and S2 ($BF_{10-MvsS2}$ = 2456000000), and moderate evidence of no difference between S1 and S2 ($BF_{10-MvsS2}$ = 0.31). Mean scores for notes forgotten were *M* = 1.6, *SD* = 6.63 for massed; *M* = 44.66, *SD* = 42.08 for S1 and *M* = 50.71, *SD* = 38.45 for S2.

At test 2.2 there was extremely strong evidence of difference in number of notes forgotten overall, ($BF_{10\ overall}$ = 375), extremely strong evidence of difference between M and S1 ($BF_{10-MvsS1}$ = 222), extremely strong evidence of difference between M and S2 ($BF_{10-MvsS2}$ = 2219), and moderate evidence of no difference between S1 and S2 ($BF_{10-MvsS2}$ = 0.294). Mean scores for notes forgotten were *M* = 1.833, *SD* = 9.30 for massed; *M* = 31.379, *SD* = 37.50 for S1 and *M* = 36.00, *SD* = 36.63 for S2 (Table 14).

Forgetting: number of pitched quarters, dotted quarters and halves, and eighth notes. At test 2.1, there was extremely strong evidence of differences in forgetting overall for quarter notes sung, ($BF_{10\ overall}$ = 103645). There was extremely strong evidence of difference in sung quarters between M and S1 ($BF_{10\text{-}MvsS1}$ = 13366); extremely strong evidence for a difference between M and S2 ($BF_{10\text{-}MvsS2}$ = 1477000) and moderate evidence of no difference in sung quarters at S1 compared to S2 ($BF_{10\text{-}MvsS1}$ = 0.307). Mean scores for quarter notes sung at test 2.1 were *M* = 31.63, *SD* = 32.22 for massed, *M* = 50.52, *SD* = 31.28 for S1 and *M* = 55.5, *SD* = 27.06 for S2 (Table 16).

At test 2.2, there was extremely strong evidence of differences in forgetting overall for quarter notes sung ($BF_{10\ overall}$ = 276); strong evidence of difference in sung quarter notes between M and S1 ($BF_{10\text{-}MvsS1}$ = 95); very strong evidence for a difference between S1and M ($BF_{10\text{-}MvsS2}$ = 1009); and moderate evidence of no difference in number of sung quarter notes at S1 compared to S2 ($BF_{10\text{-}S1vsS2}$ = 0.174). Mean scores for quarter notes sung at test 2.2 were *M* = 36.63, *SD* = 31.2 for massed, *M* = 62.31, *SD* = 20.28 for S1 and *M* = 60.89, *SD* = 20.43 for S2 (Table 16).

At test 2.1, there was extremely strong evidence of differences in forgetting overall for dotted quarter notes and half notes sung, ($BF_{10\ overall}$ = 59829). There was extremely strong evidence of difference between M and S1 ($BF_{10\text{-}MvsS1}$ = 8999); extremely strong evidence for a difference between M and S2 ($BF_{10\text{-}MvsS2}$ = 1477000) and moderate evidence of no difference at S1 compared to S2 ($BF_{10\ \text{-}MvsS1}$ = 0.30). Mean scores for dotted quarter notes and half notes sung at test 2.1 were *M* = 7.833, *SD* = 0.59 for massed, *M* = 4.069, *SD* = 3.80 for S1 and *M* = 3.571, *SD* = 3.43 for S2 (Table 17).

At test 2.2, there was extremely strong evidence of differences in forgetting overall for dotted quarter notes and half notes sung ($BF_{10\ overall}$ = 179). There was extremely strong evidence of difference between M and S1 ($BF_{10\text{-}MvsS1}$ = 153); extremely strong evidence for a difference between M and S2 ($BF_{10\text{-}MvsS2}$ = 784) and moderate evidence of no difference at S1 compared to S2 ($BF_{10\ \text{-}MvsS1}$ = 0.283). Mean scores for dotted quarter notes and half notes sung at test 2.1 were *M* = 7.767, *SD* = 0.94 for massed, *M* = 5.172, *SD* = 3.37 for S1 and *M* = 4.857, *SD* = 3.32 for S2 (Table 17).

At test 2.1, there was extremely strong evidence of differences overall in forgetting for eighth notes sung, ($BF_{10\ overall}$ = 1985000). There was extremely strong evidence of difference between M and S1($BF_{10\text{-}MvsS1}$ = 191722); extremely strong evidence for a difference between M and S2 ($BF_{10\text{-}MvsS2}$ = 11200000) and moderate evidence of no difference at S1 compared to S2 ($BF_{10\ \text{-}S1vsS2}$ = 0.302). Mean scores for eighth notes sung at test 2.1 were *M* = 11.67, *SD* = 1.16 for massed, *M* = 5.448, *SD* = 5.32 for S1 and *M* = 4.714, *SD* = 4.99 for S2 (Table 18).

At test 2.2, there was extremely strong evidence of differences overall in forgetting for eighth notes sung, ($BF_{10\ overall}$ = 22942). There was extremely strong evidence of difference between M and S1 ($BF_{10\text{-}MvsS1}$ = 4408); extremely strong evidence for a difference between M and S2 ($BF_{10\text{-}MvsS2}$ = 192736) and moderate evidence of no difference in sung quarters at S1 compared to S2 ($BF_{10\ \text{-}S1vsS2}$ = 0.313). Mean scores for eighth notes sung at test 2.2 were *M* = 11.733, *SD* = 1.11 for massed, *M* = 6.931, *SD* = 4.98 for S1 and *M* = 6.143, *SD* = 4.74 for S2 (Table 18).

Forgetting: cents off-pitch of quarter-notes, *SD* of cents off-pitch of quarter notes, *ABS* of cents off-pitch of quarter notes. There was very strong evidence for differences in forgetting overall for quarter note pitch at test 2.1 ($BF_{10\ overall}$ = 40.86); very strong evidence for difference between M and S1 ($BF_{10\text{-}MvsS1}$ = 67.74); moderate evidence for difference between M and S2 ($BF_{10\text{-}MvsS2}$ = 7.67) and no evidence of difference between S1 and S2 ($BF_{10\text{-}S1vs\ S2}$ = 0.758). Mean scores for cents off-pitch of quarter notes were *M* = 4.73, *SD* = 36.49 for massed, *M* = -162.1, *SD* = 236.3 for S1 and *M* = -66.37, *SD* = 125.8 for S2 (Table 19). *SD* of cents off-pitch of quarter notes showed extremely strong differences overall at test 2.1 ($BF_{10\ overall}$ = 1136); extremely strong evidence of difference between M and S1 ($BF_{10\text{-}Mvs1}$ = 2678); strong evidence of difference between M and S2 ($BF_{10\text{-}MvsS2}$ = 11.171); and no evidence of difference between S1 and S2 ($BF_{10\ \text{-}S1vsS2}$ = 1.477). Mean scores for *SD* of cents off-pitch of quarter notes were *M* = 34.78, *SD* = 14.55 for massed, M = 148.7, *SD* = 121.4 for S1 and *M* = 79.63, *SD* = 77.97 for S2 (Table 20). *ABS* of cents off-pitch of quarter notes showed strong differences overall at test 2.1 ($BF_{10\ overall}$ = 24.79); very strong evidence of difference between M and S1 ($BF_{10\text{-}Mvs1}$ = 37.99); strong evidence of difference between M and S2 ($BF_{10\text{-}MvsS2}$ = 14.70); and no evidence of difference between S1 and S2 ($BF_{10\ \text{-}S1vsS2}$ = 0.64). Mean scores for *ABS* of cents off-pitch of quarter notes were *M* = 36.63, *SD* = 23.26 for massed, M = 189.0, *SD* = 231.9 for S1 and *M* = 105.8, *SD* = 115.0 for S2 (Table 21).

Forgetting: Number of quarter notes off-pitch. There was evidence of no difference between the number of quarter notes off-pitch overall at test 2.1 ($BF_{10\ overall}$ = 0.119). Mean scores for number of off-pitch quarter notes at test 2.1 were *M* = 17.20, *SD* = 18.00 for massed, *M* = 17.72, *SD* = 21.10 for S1 and *M* = 0.468, *SD* = 0.341 for S2. At test 2.2 there was evidence of no difference overall ($BF_{10\ overall}$ = 0.104). Mean scores for number of off-pitch quarter notes were *M* = 21.27, *SD* = 21.13 for massed, *M* = 20.62, *SD* = 19.35 for S1 and *M* = 20.36, *SD* = 18.00 for S2 (Table 22).

Forgetting: length of quarter notes. There was no evidence of forgetting for quarter note length at either 2.1 ($BF_{10\ overall}$ = 0.371) or 2.2 ($BF_{10\ overall}$ = 0.755). There was also no evidence of forgetting for *SD* of quarter note length. (Tables 23 and 24).

Table 8

Measures of forgetting at test 2.1. Massed, 2 day, and 1 week are RIs

Measure	Massed		2 day		1 week	
	M	*SD*	*M*	*SD*	*M*	*SD*
Syllable omissions	1.7	6.64	51.21	41.52	59.29	34.62
Note omissions	1.6	6.63	44.66	42.08	50.71	38.45
Pitched quarters sung	70.63	4.97	37.28	32.95	32.5	30.04
Dot. quart. and half sung	7.833	0.59	4.069	3.80	3.571	3.43
Eighths sung	11.67	1.16	5.448	5.32	4.714	4.99

Table 9

Measures of forgetting at test 2.2. Massed, 2 day, and 1 week are RIs

Measure	Massed		2 day		1 week	
	M	*SD*	*M*	*SD*	*M*	*SD*
Syllable omissions	3.4	11.64	34.862	36.82	42.5	34.06
Note omissions	1.833	9.296	31.379	37.501	36.00	36.630
Pitched quarters sung	70.37	7.275	47.83	29.357	44.11	28.615
Dot. quart. and half sung	7.767	0.935	5.172	3.371	4.857	3.319
Eighths sung	11.733	1.112	6.931	4.985	6.143	4.743

4.3 Spacing Effects

Spacing effects: syllables omitted. For number of syllables omitted at test 3.1 there was moderate evidence for a difference overall ($BF_{10\ overall}$ = 4.467). While there was no evidence of a difference between M and S1($BF_{10\text{-}Mvs.S1}$ = 1.265), there was strong evidence of a difference at S2 compared with M ($BF_{10\text{-}MvsS2}$ = 11.936). There was no evidence of a difference in syllables omitted at S1 compared to S2 ($BF_{10\text{-}S1vsS2}$ = 0.431). Mean syllable omission scores were *M* = 55.33, *SD* = 42.76 for M; *M* =34.41, *SD* = 40.19 for S1 and *M* = 23.71, *SD* = 35.20 for S2 (Table 10) (Figure 7). For number of syllables omitted at test 3.2 there was very strong evidence of a difference overall ($BF_{10\ overall}$ = 96.26). There was strong evidence of a difference between M and S1($BF_{10\text{-}Mvs.S1}$ = 23.032) and strong evidence for a difference between M and S1($BF_{10\text{-}MvsS2}$ 31.286). There was moderate evidence for no difference in number of syllables omitted at S1 compared to S2 ($BF_{10-S1vsS2}$ = 0.276). Mean syllable omissions were *M* = 49.43 *SD* = 40.55 for M, *M* = 19.31 *SD* = 26.82 for S1 and *M* = 17.36 *SD* = 28.12 for S2 (Table 10) (Figure 8).

There was no evidence of difference overall in number of syllables sung in error in either test 3.1 ($BF_{10\ overall}$ = 0.473) or test 3.2 ($BF_{10\ overall}$ = 0.36) (Table 11). There was moderate evidence of no difference overall in number of syllables added at test 3.1 ($BF_{10\ overall}$ = 0.231) and no evidence of difference overall at test 3.2 ($BF_{10\ overall}$ = 0.644) (Table 12). There was no evidence of difference in the number of trials required to relearn to song to criterion at the third session ($BF_{10\ overall}$ = 1.113) (Table 11). Mean number of trials to criterion were *M* = 3.37, *SD* = 1.43 for massed, *M* = 2.72, *SD* = 1.41 for S1 and *M* = 2.46, *SD* = 1.48 for S2.

Table 10

Syllable omissions by test. Massed, 2 day, and 1 week ISIs.

Measure	Massed *M* (*SD*)	2 day *M* (*SD*)	1 week *M* (*SD*)	BF_{10} Overall	BF_{10} M vs. S1	BF_{10} M vs. S2	BF_{10} S1 vs. S2
Syllable omissions: test 1.3	0.033 (0.183)	0.069 (0.258)	0.036 (0.189)	0.125	0.310	0.266	0.305
Syllable omissions: test 2.1	1.7 (6.64)	51.21 (41.52)	59.29 (34.62)	41610000	372600	2456000000	0.349
Syllable omissions: test 2.2	3.4 (11.64)	34.86 (36.82)	42.5 (34.06)	4557	499	58228	0.353
Syllable omissions: test 2.4	3.4 (14.79)	1.39 (4.97)	1.43 (4.20)	0.145	0.324	0.323	0.27
Syllable omissions: test 3.1	55.33 (42.76)	34.41 (40.19)	23.71 (35.20)	4.467	1.265	11.936	0.431
Syllable omissions: test 3.2	49.43 (40.55)	19.31 (26.82)	17.36 (28.12)	96.26	23.032	31.286	0.276

Note. 2.1 and 2.2 are the first tests in session 2, which show forgetting.

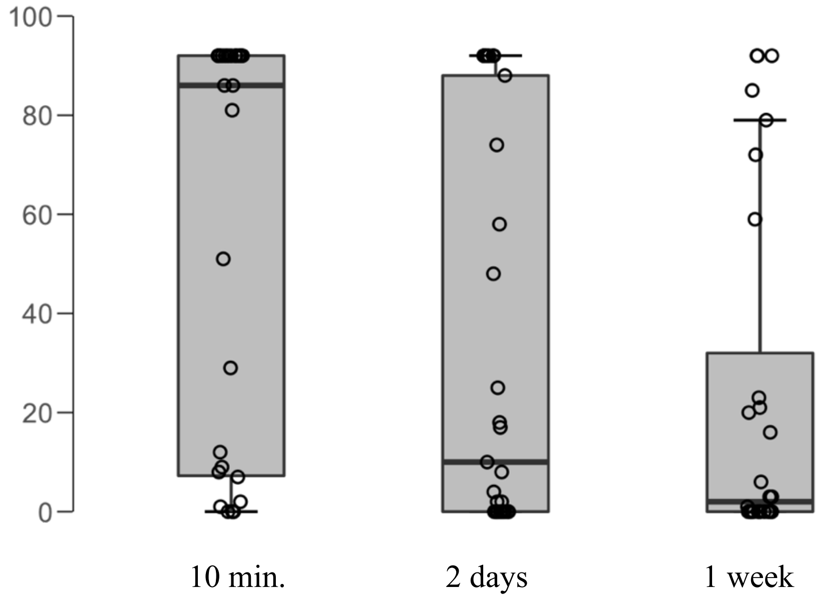

Figure 7. Syllable omissions cued by first notes (test 3.1)

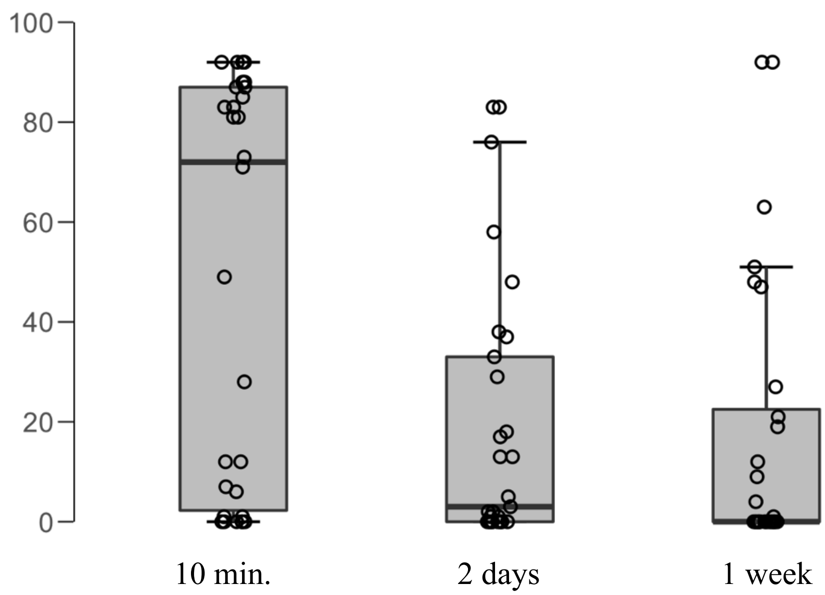

Figure 8. Syllable omissions cued by first notes and words (test 3.2)

Table 11

Syllable errors by test. Massed, 2 day, and 1 week are ISIs.

Measure	Massed *M* (*SD*)	2 day *M* (*SD*)	1 week *M* (*SD*)	BF_{10} Overall	BF_{10} Massed vs. 2 day	BF_{10} Massed vs. 1 week	BF_{10} 2 day vs. 1 week
Syllable errors: test 1.3	0.433 (0.65)	0.500 (0.668)	0.625 (0.99)	0.145	0.165	0.215	0.179
Syllable errors: test 2.1	0.87 (1.51)	1.98 (3.19)	2.02 (3.38)	0.362	0.909	0.871	0.268
Syllable errors: test 2.2	1.567 (2.239)	3.672 (4.591)	2.786 (3.573)	0.767	2.113	0.738	0.353
Syllable errors: test 2.4	0.58 (1.13)	1.3 (2.01)	1.23 (1.79)	0.376	0.876	0.834	0.272
Syllable errors: test 3.1	.92 (1.36)	2.52 (4.3)	2.32 (3.29)	0.473	0.981	1.144	0.272
Syllable errors: test 3.2	1.57 (3.09)	3.12 (4.2)	2.43 (2.51)	0.36	0.79	0.467	0.339

Note. 2.1 and 2.2 are the first tests in session 2, which show forgetting.

Table 12

Syllables added by test. Massed, 2 day, and 1 week are ISIs.

Measure	Massed *M* (*SD*)	2 day *M* (*SD*)	1 week *M* (*SD*)	BF_{10} Overall	BF_{10} Massed vs. 2 day	BF_{10} Massed vs. 1 week	BF_{10} 2 day vs. 1 week
Syllables added: test 2.1	0	0	0				
Syllables added: test 2.1	0.1 (0.31)	0.14 (0.35)	0.07 (0.26)	0.133	0.287	0.283	0.35
Syllables added: test 2.2	0.033 (0.183)	).069 (0.258)	0.107 (0.315)	0.164	0.310	0.441	0.298
Syllables added: test 2.4	0	0	0				
Syllables added: test 3.1	0.1 (0.31)	0.03 (0.19)	0.21 (0.42)	0.231	0.506	0.631	0.276
Syllables added: test 3.2	0.03 (0.18)	0.14 (0.35)	0.14 (0.36)	0.644	0.399	0.483	1.664

Note. 2.1 and 2.2 are the first tests in session 2, which show forgetting.

Table 13

Session 3 trials to criterion. Massed, 2 day, and 1 week ISIs.

Measure	Massed *M* (*SD*)	2 day *M* (*SD*)	1 week *M* (*SD*)	BF_{10} Overall	BF_{10} Massed vs. 2 day	BF_{10} Massed vs. 1 week	BF_{10} 2 day vs. 1 week
Session 3 trials to criterion	3.37 (1.45)	2.72 (1.41)	2.46 (1.48)	1.113	0.925	2.616	0.324

Spacing effect: number of sung notes omitted. At test 3.1, there was moderate evidence of a difference overall for number of sung notes omitted, ($BF_{10\ overall}$ = 4.896). There was no evidence of difference in sung notes omitted between M and S1($BF_{10\text{-}MvsS1}$ = 2.171); strong evidence for a difference between M and S2 ($BF_{10\text{-}MvsS2}$ = 10.141) and moderate evidence of no difference in sung notes omitted at S1 compared to S2 ($BF_{10\text{-}S1vsS2}$ = 0.318). At test 3.1 mean note omission scores were *M* = 51.47, *SD* = 41.36 for massed; *M* =27.48, *SD* = 9.89 for S1 and *M* = 21.11, *SD* = 34.79 for S2 participants (Table 14) (Figure 9).

At test 3.2, there was extremely strong evidence of a difference overall for number of sung notes omitted ($BF_{10\ overall}$ = 205); strong evidence of difference in sung notes omitted between M and S1($BF_{10\text{-}MvsS1}$= 53.09); strong evidence for a difference between S2 and M ($BF_{10\text{-}MvsS2}$ = 33.47); and moderate evidence of no difference in number of notes omitted at S1 compared to S2 ($BF_{10\text{-}S1\ vsS2}$ = 0.271). At test 3.2 mean note omission scores were *M* = 45.33, *SD* = 40.17 for M; *M* = 12.79, *SD* = 26.13 for S1 and *M* = 13.89, *SD* = 26.21 for S2 (Table 14) (Figure 10).

Table 14

Note omissions by test. Massed, 2 day, and 1 week are ISIs

Measure	Massed *M* (*SD*)	2 day *M* (*SD*)	1 week *M* (*SD*)	BF_{10} Overall	BF_{10} Massed vs. 2 day	BF_{10} Massed vs. 1 week	BF_{10} 2 day vs. 1 week
Notes omitted test 1.1	1.3 (3.25)	1.28 (3.41)	1.14 (2.99)	0.10	0.26	0.27	0.27
Notes omitted test 2.1	1.6 (6.63)	44.66 (42.08)	50.71 (38.45)	125369.25	15987.63	166660000	0.31
Notes omitted test 2.2	1.833 (9.296)	31.379 (37.501)	36.00 (36.630)	375.572	222.598	2219.031	0.294
Notes omitted test 2.4	3.4 (14.79)	1.21 (4.54)	0.54 (1.84)	0.19	0.34	0.41	0.34
Notes omitted test 3.1	51.47 (41.36)	27.48 (39.89)	21.11 (34.79)	4.90	2.17	10.14	0.318
Notes omitted test 3.2	45.33 (40.17)	12.79 (26.13)	13.89 (26.21)	205.94	53.09	33.47	0.27

Note. 2.1 and 2.2 are the first tests in session 2, which show forgetting

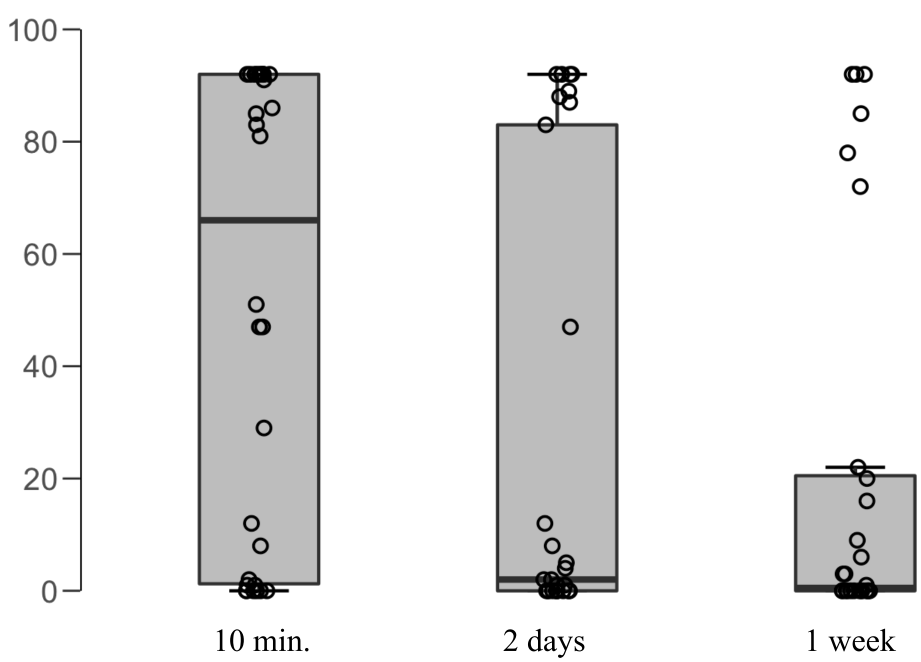

Figure 9. Note omissions (test 3.1) cued by first notes

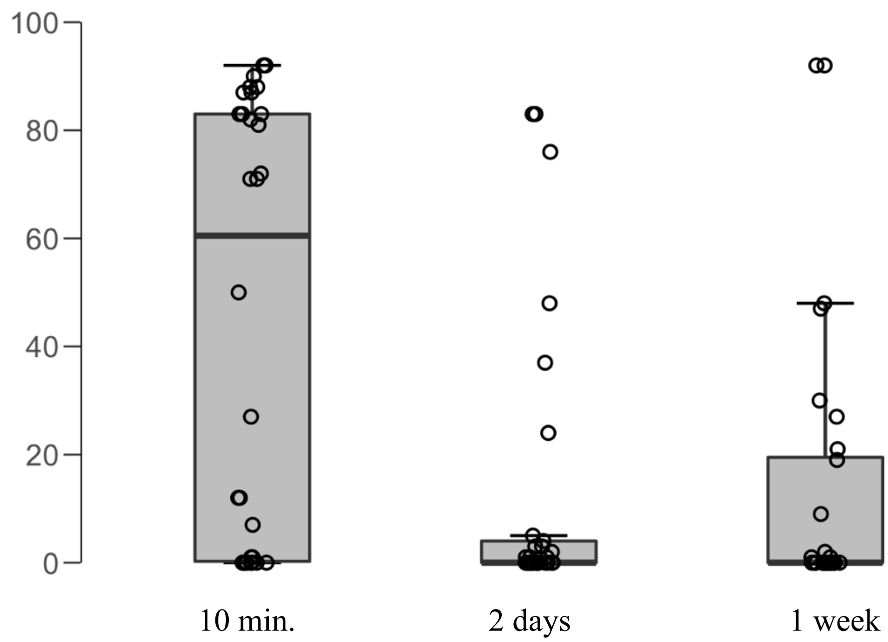

Figure 10. Note omissions (test 3.2) cued by first notes and words

A comparison of syllables omitted and notes omitted in the final session tests shows that consistently more notes were produced than words, with mean scores and spacing effects exactly parallel (Table 15) (Figures 21-24).

Table 15

Comparison of syllable and note omissions, Tests 3.1 and 3.2. Massed, 2 day, and 1 week are ISIs

Measure	Massed M (SD)	2 day *M* (*SD*)	1 week *M* (*SD*)	BF_{10} Overall	BF_{10} Massed vs. 2 day	BF_{10} Massed vs. 1 week	BF_{10} 2 day vs. 1 week
Syllable omissions: test 3.1	55.33 (42.76)	34.41 (40.19)	23.71 (35.20)	4.467	1.265	11.936	0.431
Syllable omissions: test 3.2	49.43 (40.55)	19.31 (26.82)	17.36 (28.12)	96.26	23.032	31.286	0.276
Note omissions test 3.1	51.47 (41.36)	27.48 (9.89)	21.11 (34.79)	4.90	2.17	10.14	0.318
Note omissions test 3.2	45.33 (40.17)	12.79 (26.13)	13.89 (26.21)	205.94	53.09	33.47	0.27

Spacing effects: number of sung quarters, dotted quarters and halves, and eighth notes. The same differences which showed up in notes omitted also appeared in the total numbers of notes of each type that were sung. At test 3.1, there was moderate evidence of a difference overall for quarter notes sung, ($BF_{10\ overall}$ = 5.19). There was no evidence of difference in sung notes omitted at S1 compared to M ($BF_{10\text{-}MvsS1}$ = 2.239); strong evidence for a difference at S1compared to M ($BF_{10\text{-}MvsS2}$ = 10.857) and moderate evidence of no difference in sung notes omitted at S1 compared to S2 ($BF_{10\text{-}MvsS1}$ = 0.318). Mean scores for quarter notes sung at test 3.1 were *M* = 31.63, *SD* = 32.22 for massed, *M* = 50.52, *SD* = 31.28 for S1 and *M* = 55.5, *SD* = 27.06 for S2 (Table 16) (Figure 11). At test 3.2, there was extremely strong evidence of a difference overall for quarter notes sung ($BF_{10\ overall}$ = 222.108); strong evidence of difference in sung quarter notes between M and S1 ($BF_{10\text{-}MvsS1}$ = 62.27); strong evidence for a difference between S2 and M ($BF_{10\text{-}MvsS2}$ = 31.241); and moderate evidence of no difference between S1 and S2 ($BF_{10=S1vsS2}$ = 0.276). Mean scores for quarter notes sung at test 3.2 were *M* = 36.63, *SD* = 31.20 for massed, *M* = 62.31, *SD* = 20.28 for S1 and *M* = 60.89, *SD* = 20.43 for S2 (Table 16) (Figure 12).

At test 3.1, there was moderate evidence of a difference overall for dotted quarter and half notes sung, ($BF_{10\ overall}$ = 6.59). There was no evidence of difference between M and S1 ($BF_{10\ \text{-}MvsS1}$ = 2.384); strong evidence for a difference between M and S2 ($BF_{10\text{-}MvsS2}$ = 14.726) and moderate evidence of no difference for S1 compared to S2 ($BF_{10\text{-}S1vsS2}$ = 0.327). Mean scores for dotted quarter and half notes at test 3.1 were *M* = 3.467, *SD* = 3.60 for M, *M* = 5.621, *SD* = 3.54 for S1 and *M* = 6.214, *SD* = 2.94 for S2 (Table 17) (Figure 13). At test 3.2, there was extremely strong evidence of a difference overall for dotted quarter and half notes sung ($BF_{10\ overall}$ = 117); strong evidence of difference between M and S1($BF_{10\ MvsS1}$ = 48.729); strong evidence for a difference between S2 and M ($BF_{10\text{-}MvsS2}$ = 16.352); and moderate evidence of no difference between S1 and S2 ($BF_{10\text{-}S1vsS2}$ = 0.294). Mean scores for dotted quarter and half notes at test 3.2 were *M* = 4.067, *SD* = 3.58 for massed, *M* = 6.931, *SD* = 2.30 for S1 and *M* = 6.643, *SD* = 2.35 for S2 (Table 17) (Figure 14).

At test 3.1, there was moderate evidence of a difference overall for eighth notes sung ($BF_{10\ overall}$ = 3.228). There was no evidence of difference for S1 compared to M ($BF_{10\ \text{-}MvsS1}$ = 1.829); moderate evidence for a difference between S2 and M ($BF_{10\text{-}MvsS2}$ = 6.785) and moderate evidence of no difference between S1 and S2 ($BF_{10\text{-}S1vsS2}$ = 0.31). Mean scores for eighths sung at test 3.1 were *M* = 4.9, *SD* = 5.36 for massed, *M* = 7.897, *SD* = 5.25 for S1 and *M* = 8.69, *SD* = 4.75 for S2 (Table 18) (Figure 15). At test 3.2, there was strong evidence of a difference overall for eighth notes sung ($BF_{10\ overall}$ = 92.183); strong evidence of difference between S1 and M ($BF_{10\ \text{-}MvsS1}$ = 24.065); strong evidence for a difference between S2 and M ($BF_{10\text{-}MvsS2}$ = 16.352); and moderate evidence of no difference between S1 and S2 ($BF_{10\text{-}S1vsS2}$ = 0.294). Mean scores at test 3.2 were *M* = 5.5, *SD* = 5.44 for massed, *M* = 9.621, *SD* = 3.77 for S1 and *M* = 9.75, *SD* = 3.69 for S2 (Table 18) (Figure 16). No notes were added to the 94 notes of the song in any tests under any conditions.

Table 16

Number of sung quarter notes by test. Massed, 2 day, and 1 week are ISIs

Measure	Massed *M* (*SD*)	2 day *M* (*SD*)	1 week *M* (*SD*)	BF_{10} Overall	BF_{10} Massed vs. 2 day	BF_{10} Massed vs. 1 week	BF_{10} 2 day vs. 1 week
S. quart Test 1.1	71.23 (1.91)	70.97 (2.15)	70.89 (3.35)	0.115	0.294	0.293	0.269
S. quart test 2.1	70.63 (4.97)	37.28 (32.95)	32.5 (30.04)	103645.44	13366.045	1477000	0.307
S. quart test 2.2	70.37 (7.275)	47.83 (29.357)	44.11 (28.615)	276.539	95.916	1009.789	0.174
S. quart test 2.4	69.17 (11.64)	70.93 (3.45)	70.96 (2.73)	0.163	0.341	0.347	0.27
S. quart test 3.1	31.63 (32.22)	50.52 (31.28)	55.5 (27.06)	5.19	2.239	10.857	0.318
S. quart test 3.2	36.63 (31.2)	62.31 (20.28)	60.89 (20.43)	222.108	62.27	31.241	0.276

Note. 2.1 and 2.2 are the first tests in session 2, which show forgetting.

Table 17

Number of pitched dotted quarter and half notes by test. Massed, 2 day, and 1 week are ISIs

Measure	Massed *M* (*SD*)	2 day *M* (*SD*)	1 week *M* (*SD*)	BF_{10} Overall	BF_{10} Massed vs. 2 day	BF_{10} Massed vs. 1 week	BF_{10} 2 day vs. 1 week
D. quarters and half n. Test 1.1	7.667 (.80)	7.828 (0.54)	7.929 (0.38)	0.305	0.371	0.741	0.354
D. quarters and half n. test 2.1	7.833 (0.59)	4.069 (3.80)	3.571 (3.43)	59829	8999	881679	59829
D. quarters and half n. test 2.2	7.767 (0.935)	5.172 (3.371)	4.857 (3.319)	179	153	784.	0.283
D. quarters and half n. test 2.4	7.567 (1.33)	7.893 (0.42)	7.893 (0.31)	0.328	0.506	0.518	0.27
D. quarters and half n. test 3.1	3.467 (3.60)	5.621 (3.54)	6.214 (2.94)	6.59	2.384	14.726	0.327
D. quarters and half n. test 3.2	4.067 (3.58)	6.931 (2.30)	6.643 (2.35)	117.079	48.729	16.352	0.294

Note. 2.1 and 2.2 are the first tests in session 2, which show forgetting.

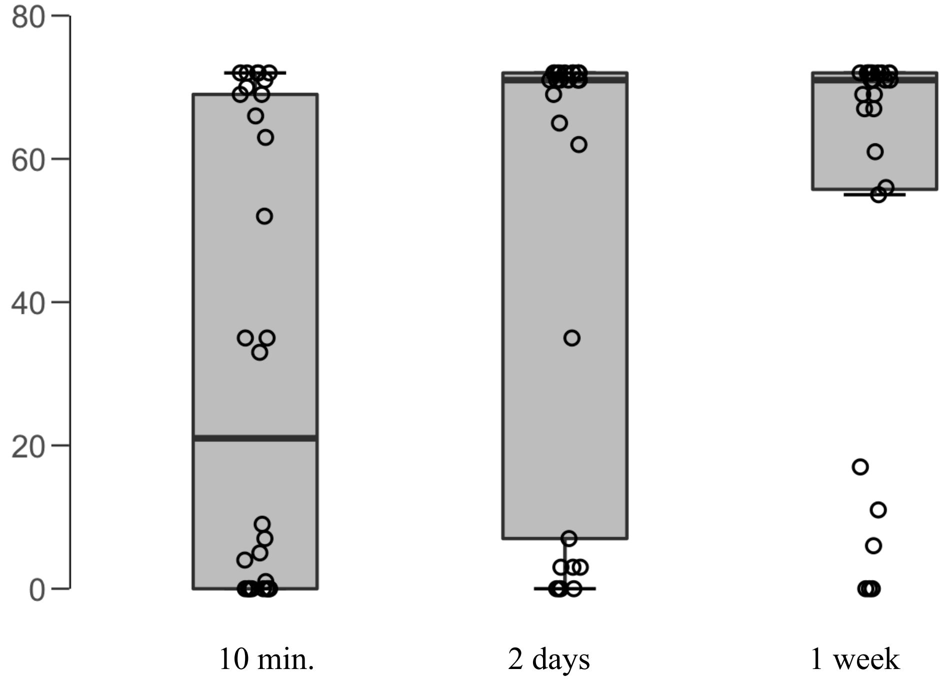

Figure 11. Number of sung quarter notes, test 3.1

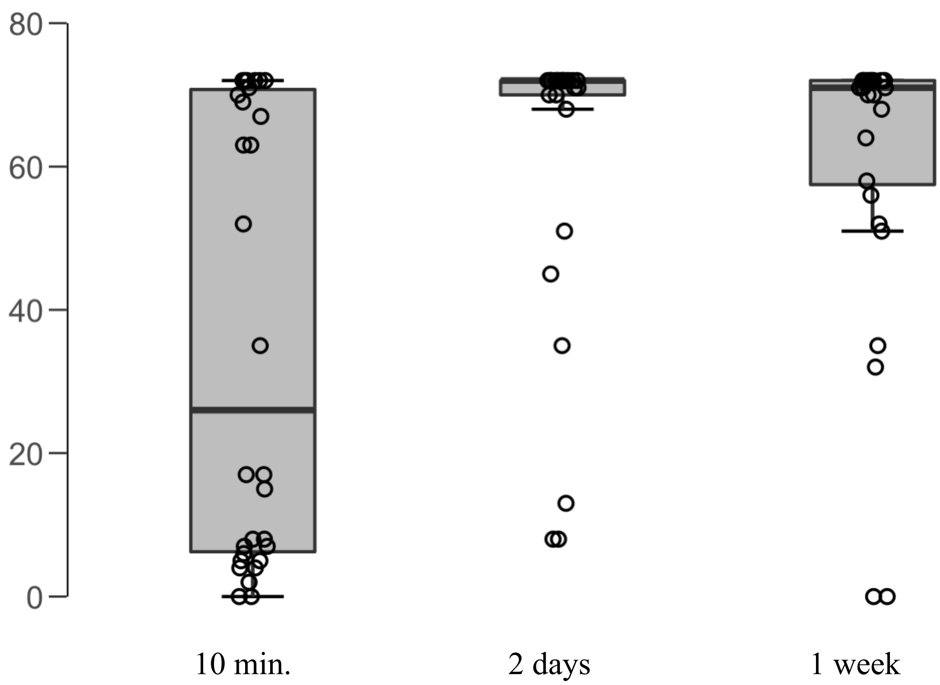

Figure 12. Number of sung quarter notes, test 3.2.

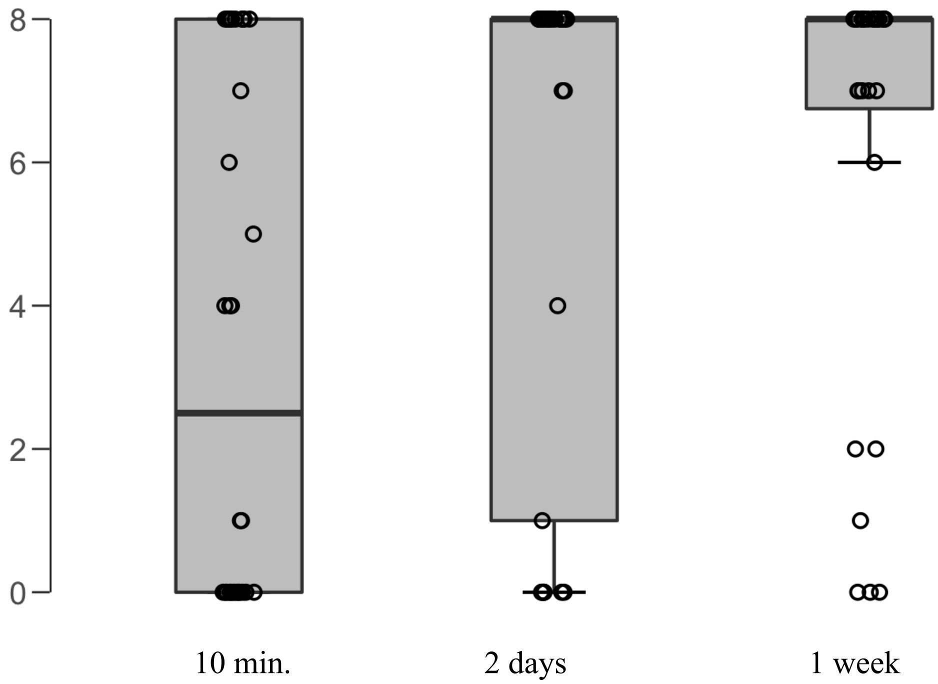

Figure 13. Number of dotted quarters and half notes, test 3.1.

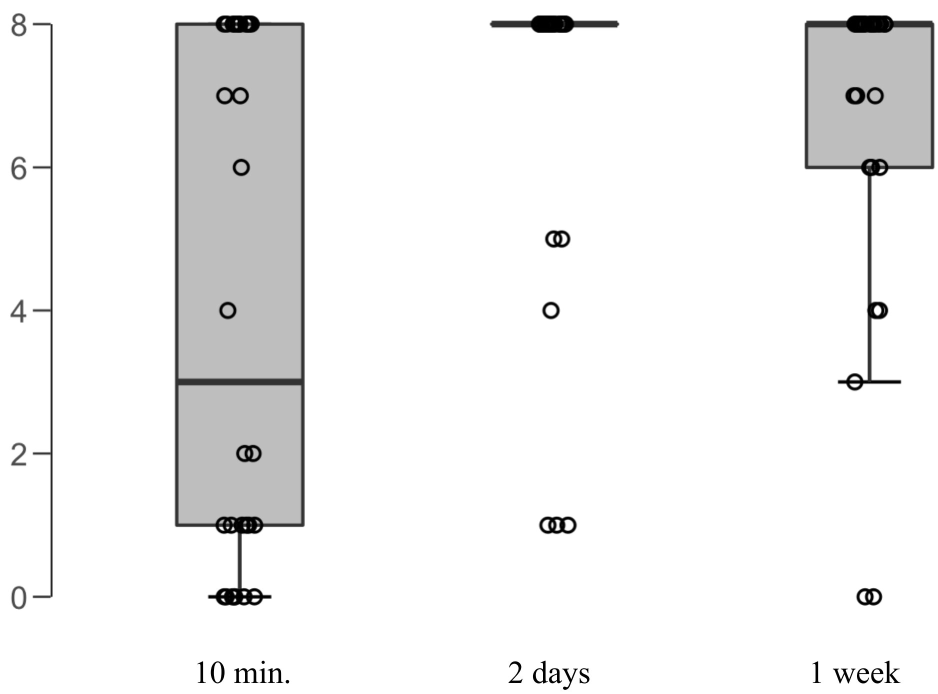

Figure 14. Number of sung dotted quarters and half notes, test 3.2.

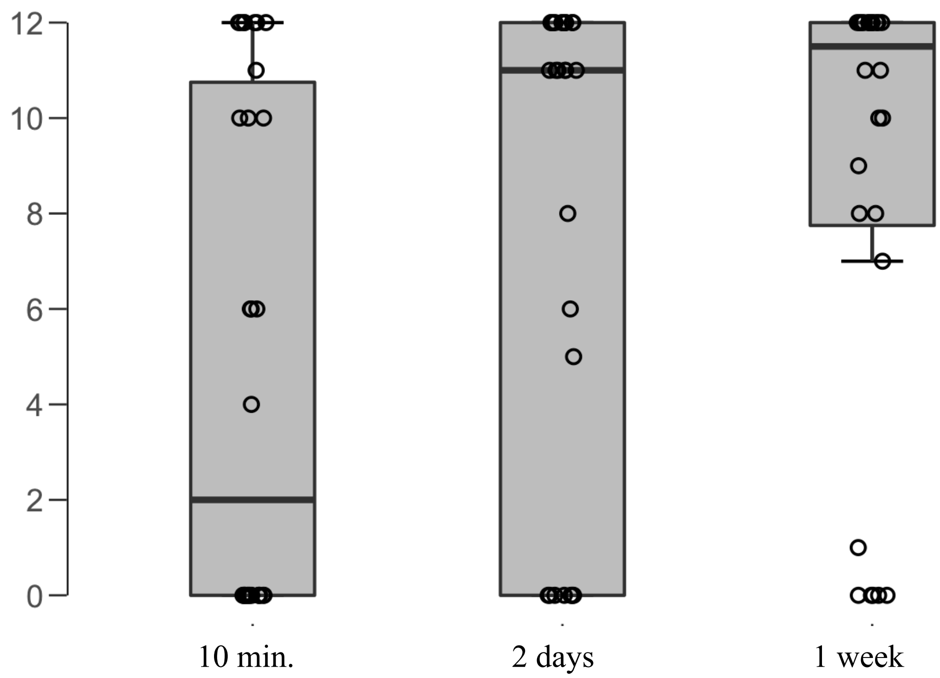

Figure 15. Number of sung eighth notes, test 3.1.

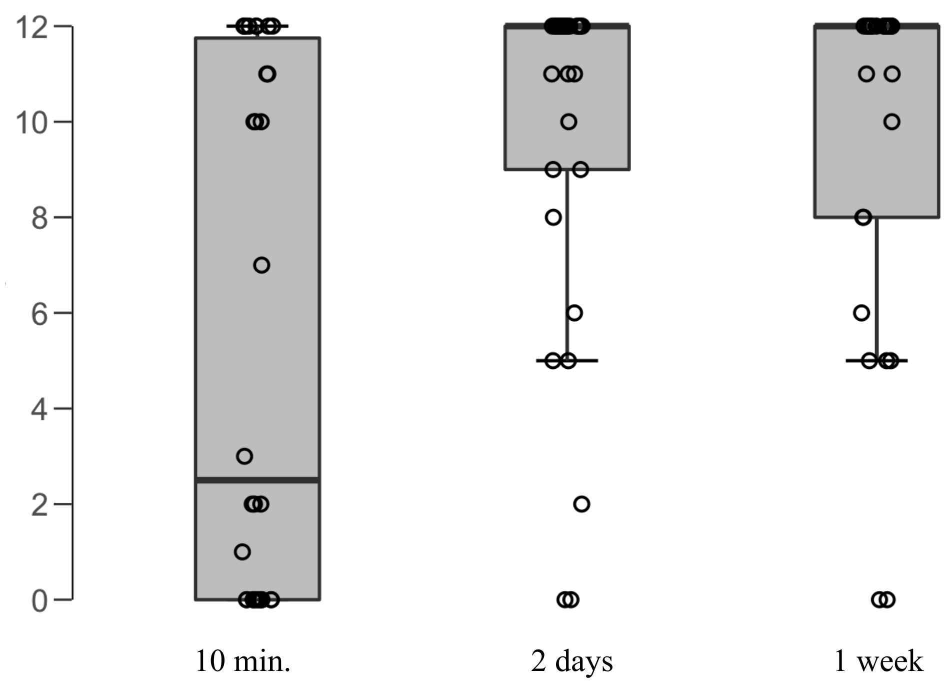

Figure 16. Number of sung eighth notes, test 3.2.

Table 18

Number of pitched eighth notes by test. Massed, 2 day, and 1 week are ISIs

Measure	Massed *M* (*SD*)	2 day *M* (*SD*)	1 week *M* (*SD*)	BF_{10} Overall	BF_{10} Massed vs. 2 day	BF_{10} Massed vs. 1 week	BF_{10} 2 day vs. 1 week
P. eighth Test 1.1	11.67 (1.21)	11.76 (1.3)	11.82 0.39)	0.116	0.273	0.317	0.275
P. eighth test 2.1	11.67 (1.16)	5.448 (5.32)	4.714 (4.99)	1985000	191722	11200000	0.302
P. eighth test 2.2	11.733 (1.112)	6.931 (4.985)	6.143 (4.743)	22942	4408	192736	0.313
P. eighth test 2.4	11.57 (1.85)	11.64 (0.87)	11.68 (0.90)	0.108	0.271	0.276	0.272
P. eighth test 3.1	4.9 (5.36)	7.897 (5.25)	8.679 4.75)	3.228	1.829	6.785	0.31
P. eighth test 3.2	5.5 (5.44)	9.621 (3.77)	9.75 (3.69)	92.18	24.06	29.87	0.27

Note. 2.1 and 2.2 are the first tests in session 2, which show forgetting.

Spacing effects: cents off-pitch quarter-notes, *SD* of cents off-pitch quarter notes, *ABS* of cents off-pitch quarter notes . At test 3.1 there was no evidence of difference overall for cents off-pitch of quarter notes ($BF_{10\ overall}$ = 0.786), and at test 3.2 there was moderate evidence of no difference overall for cents off-pitch of quarter notes ($BF_{10\ overall}$ = 0.121) (Table 19) (Figures 17 and 18). For SD of quarter notes off-pitch, at test 3.1, there was no evidence of difference overall ($BF_{10\ overall}$ = 0.79). At test 3.2, there was moderate evidence of no difference overall ($BF_{10\ overall}$ = 0.12) (Table 20). For *ABS* of quarter notes off-pitch, at test 3.1, there was no evidence of difference overall ($BF_{10\ overall}$ = 0.439). At test 3.2, there was moderate evidence of no difference overall ($BF_{10\ overall}$ = 0.157) (Table 21).

Spacing effects; Number of quarter notes off-pitch. At test 3.1 there was no evidence difference overall for number of quarter notes more than 50 cents off-pitch ($BF_{10\ overall}$ = 0.746). At test 3.2 there was also no evidence of difference overall ($BF_{10\ overall}$ = 0.496) (Table 22).

Table 19

Cents off-pitch for quarter notes. Massed, 2 day, and 1 week are ISIs

Measure	Massed *M* (*SD*)	2 day *M* (*SD*)	1 week *M* (*SD*)	BF_{10} Overall	BF_{10} Massed vs. 2 day	BF_{10} Massed vs. 1 week	BF_{10} 2 day vs. 1 week
Cents o.p. quarters Test 1.1	16.21 (31.68)	4.91 (60.62)	-1.16 (22.93)	0.285	0.372	2.67	0.297
Cents o.p. quarters Test 1.3	18.67 (36.81)	17.418 (31.65)	9.803 (22.77)	0.174	0.266	0.439	0.420
Cents o.p. quarters test 2.1	4.73 (36.49)	-162.1 (236.3)	-66.37 (125.8)	40.86	67.74	7.68	0.758
Cents o.p. quarters test 2.4	25.66 (42.37)	3.9 (39.88)	4.62 (43.3)	0.762	1.40	1.13	0.27
Cents o.p. quarters test 3.1	-84.64 (172)	-17.3 (57.43)	-20.1 (73.7)	0.786	0.991	0.923	0.293
Cents o.p. quarters test 3.2	-26.09 (77.16)	-33.07 (129.7)	-19.94 (53.79)	0.121	0.281	0.293	0.3

Note. 2.1 is the first test in session 2, which shows forgetting.

Table 20

SD of cents off-pitch for quarter notes. Massed, 2 day, and 1 week are ISIs

Measure	Massed *M* (*SD*)	2 day *M* (*SD*)	1 week *M* (*SD*)	BF_{10} Overall	BF_{10} Massed vs. 2 day	BF_{10} Massed vs. 1 week	BF_{10} 2 day vs. 1 week
SD cents o.p.quart. Test 1.1	52.16 (84.12)	45.92 (49.48)	46.9 (57.96)	0.109	0.278	0.275	0.269
SD cents o.p.quart. Test 1.3	31.69 (18.16)	38.53 (18.58)	35.96 (20.37)	0.220	0.618	0.358	0.297
SD cents o.p.quart. test 2.1	34.78 (14.55)	148.7 (121.4)	79.63 (77.97)	1136	2678	11.17	1.48
SD cents o.p.quart. test 2.4	35.84 (28.38)	46.54 (48.88)	34.34 (11.04)	0.251	0.414	0.274	0.537
SD cents o.p.quart. test 3.1	74.88 (55.43)	61.69 (49.87)	64.84 (66.2)	0.156	0.397	0.339	0.294
SD cents o.p.quart. test 3.2	93.56 (110.8)	77.26 (99.66)	68.71 (78.47)	0.152	0.314	0.399	0.287

Note. 2.1 and 2.2 are the first tests in session 2, which show forgetting.

Table 21

ABS of cents off-pitch for quarter notes. Massed, 2 day, and 1 week are ISIs

Measure	Massed *M* (*SD*)	2 day *M* (*SD*)	1 week *M* (*SD*)	BF_{10} Overall	BF_{10} Massed vs. 2 day	BF_{10} Massed vs. 1 week	BF_{10} 2 day vs. 1 week
ABS cents o.p.quart. Test 1.1	47.98 (66.19)	44.01 (49.60)	38.72 (38.77)	0.122	0.272	0.317	0.291
ABS cents o.p.quart. Test 1.3	38.69 (26.30)	38.66 (21.65)	32.41 (12.23)	0.198	0.264	0.463	0.562
ABS cents o.p.quart. test 2.1	36.63 (23.26)	189.0 (231.9)	105.8 (115.0)	25.79	37.99	14.70	0.643
ABS cents o.p.quart. test 2.4	44.71 (34.03)	43.51 (26.54)	41.37 (25.70)	0.11	0.269	0.287	0.281
ABS cents o.p.quart. test 3.1	113.1 (162.4)	61.17 (46.44)	69.05 (50.75)	0.439	0.693	0.579	0.329
ABS cents o.p.quart. test 3.2	93.42 (98.77)	82.60 (124.8)	67.13 (49.29)	0.157	0.170	0.300	0.185

Note. 2.1 and 2.2 are the first tests in session 2, which show forgetting.

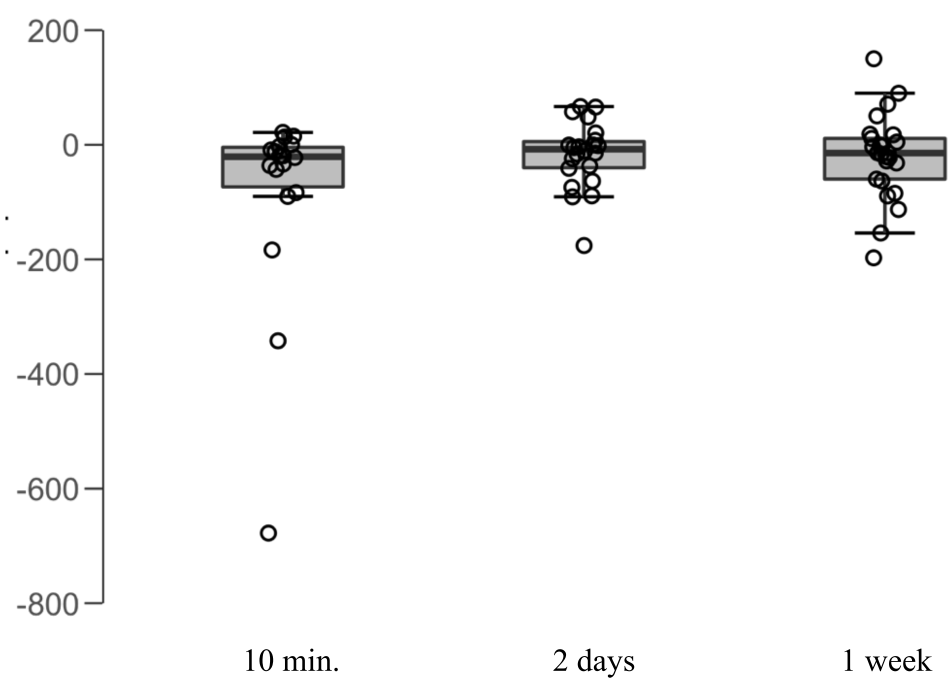

Figure 17. Test 3.1 Mean cents off-pitch quarters.

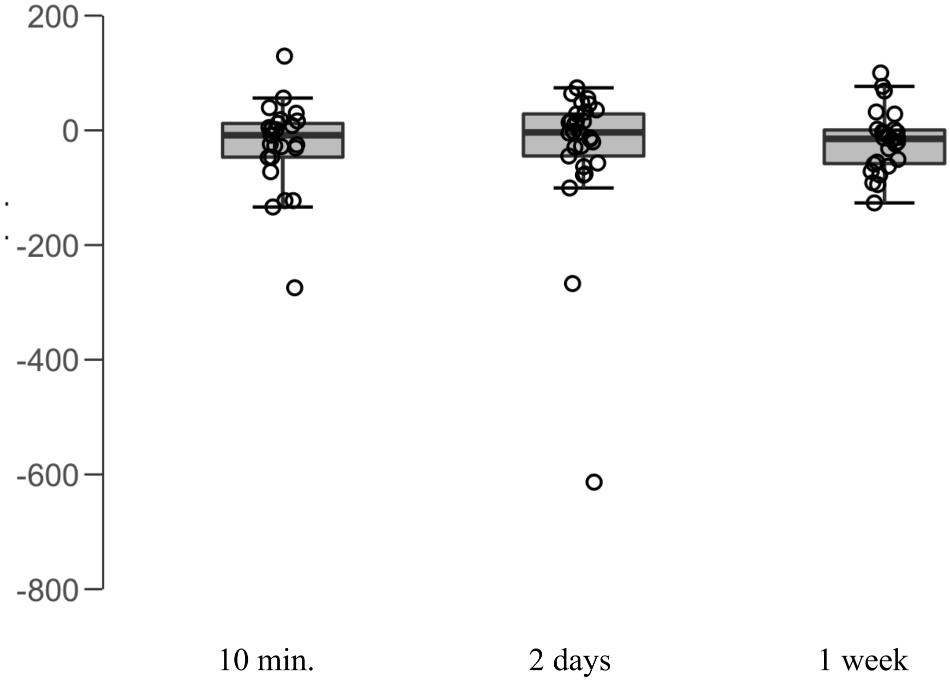

Figure 18. Test 3.2. Mean cents off-pitch quarters.

Table 22

Number of off-pitch quarter notes. Massed, 2 day, and 1 week are ISIs

Measure	Massed *M* (*SD*)	2 day *M* (*SD*)	1 week *M* (*SD*)	BF_{10} Overall	BF_{10} Massed vs. 2 day	BF_{10} Massed vs. 1 week	BF_{10} 2 day vs. 1 week
Num. o.p. quarters Test 1.1	15.37 (19.54)	15.55 (15.30)	13.00 (12.98)	0.121	0.264	0.301	0.325
Num. o.p. quarters Test 1.3	19.20 (19.84)	19.07 (15.82)	14.18 (8.70)	0.216	0.264	0.501	0.632
Num. o.p. quarters test 2.1	17.20 (18.00)	17.72 (21.10)	14.75 (18.96)	0.119	0.156	0.174	0.179
Num. o.p. quarters test 2.2	21.27 (21.13)	20.62 (19.35)	20.36 (18.00)	0.104	0.266	0.270	0.268
Num. o.p. quarters test 2.4	21.00 (22.47)	21.54 (16.16)	20.61 (18.53)	0.104	0.267	0.267	0.274
Num. o.p. quarters test 3.1	12.43 (17.11)	21.28 (20.67)	23.25 (20.93)	0.746	1	1.806	0.283
Num. o.p. quarters test 3.2	17.23 (20.76)	23.62 (16.36)	27.25 (20.43)	0.496	0.54	1.093	0.337

Note. 2.1 and 2.2 are the first tests in session 2, which show forgetting.

Table 23

Mean quarter note length in seconds. Massed, 2 day, and 1 week are ISIs

Measure	Massed *M* (*SD*)	2 day *M* (*SD*)	1 week *M* (*SD*)	BF_{10} Overall	BF_{10} Massed vs. 2 day	BF_{10} Massed vs. 1 week	BF_{10} 2 day vs. 1 week
Quarter length Test 1.1	0.42 (0.02)	0.42 (0.01)	0.417 (0.02)	0.115	0.281	0.269	0.31
Quarter length test 2.1	0.41 (0.03)	0.39 (0.08)	0.379 0.06)	0.371	0.436	1.865	0.35
Quarter length test 2.2	0.408 (0.027)	0.392 (0.051)	0.378 (0.064)	0.755	0.681	2.167	0.371
Quarter length test 2.4	0.41 (0.03)	0.4 (0.02)	0.4 (0.03)	0.819	1.25	1.313	0.275
Quarter length test 3.1	0.38 (0.04)	0.39 (0.0.03)	0.37 (0.04)	0.257	0.42	0.323	0.622
Quarter length test 3.2	0.39 (0.06)	0.39 (0.03)	0.38 (0.03)	0.151	0.3	0.363	0.319

Note. 2.1 and 2.2 are the first tests in session 2, which show forgetting.

Table 24

SD of quarter note length in seconds. Massed, 2 day, and 1 week are ISIs

Measure	Massed M (SD)	2 day *M* (*SD*)	1 week *M* (*SD*)	BF_{10} Overall	BF_{10} Massed vs. 2 day	BF_{10} Massed vs. 1 week	BF_{10} 2 day vs. 1 week
SD quart. length Test 1.1	0.04 (0.01)	0.032 (0.01)	0.036 (0.01)	0.216	0.482	0.266	0.669
SD quart. length test 2.1	0.40 (0.42)	0.60 (0.51)	0.56 (0.44)	0.439	0.865	1.065	0.324
SD quart. length test 2.2	0.071 (0.011)	0.090 (0.043)	0.089 (0.029)	1.630	2.172	11.84	0.288
SD quart. length test 2.4	0.14 (0.42)	0.07 (0.01)	0.0768 (0.01)	0.21	0.398	0.379	0.588
SD quart. length test 3.1	0.08 (0.42)	0.08 (0.01)	0.08 (0.02)	0.138	0.344	0.308	0.302
SD quart. length test 3.2	0.09 (0.05)	0.08 (0.02)	0.08 (0.02)	0.142	0.338	0.313	0.283

Note. 2.1 and 2.2 are the first tests in session 2, which show forgetting.

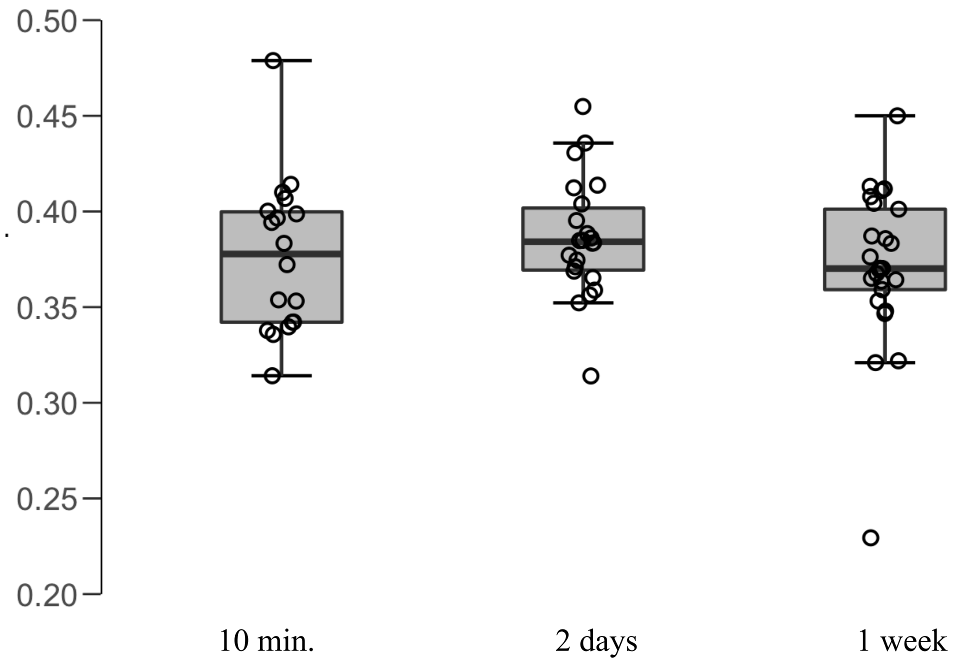

Figure 19. Test 3.1 Mean length of quarter notes in seconds.

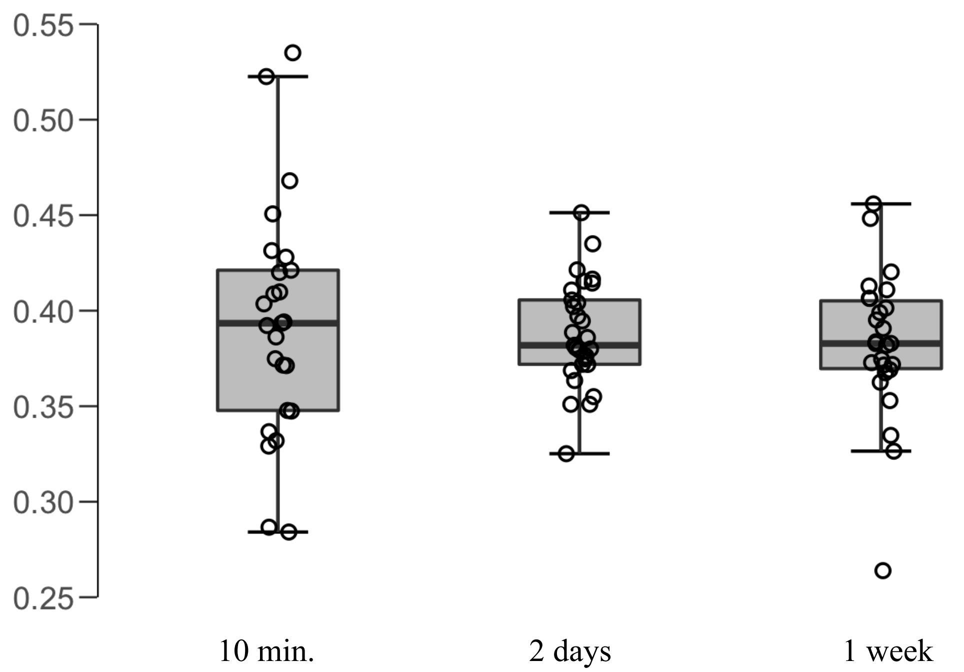

Figure 20. Test 3.2 Mean length of quarter notes in seconds.

Spacing effects: Length of quarter notes. There was moderate evidence for no difference in quarter note length at test 3.1 ($BF_{10\ overall}$ = 0.257) or at test 3.2 ($BF_{10\ overall}$ = 0.151) (Table 22) (Figures 19 and 20). For *SD* of quarter note length there was moderate evidence of no difference overall at test 3.1 ($BF_{10\ overall}$ = 0.138) or at test 3.2 ($BF_{10\ overall}$ = 0.142) (Table 23).

Correlation: Syllables omitted and notes omitted. A Bayesian Pearson correlation showed a strong correlation between syllables and notes omitted at test 2.1 (*r* = 0.940), at test 2.2 (*r* = 0.967), at test 3.1 (*r* = 0.955*)* and at test 3.2 (*r* = 0.947) (Figures 21-24).

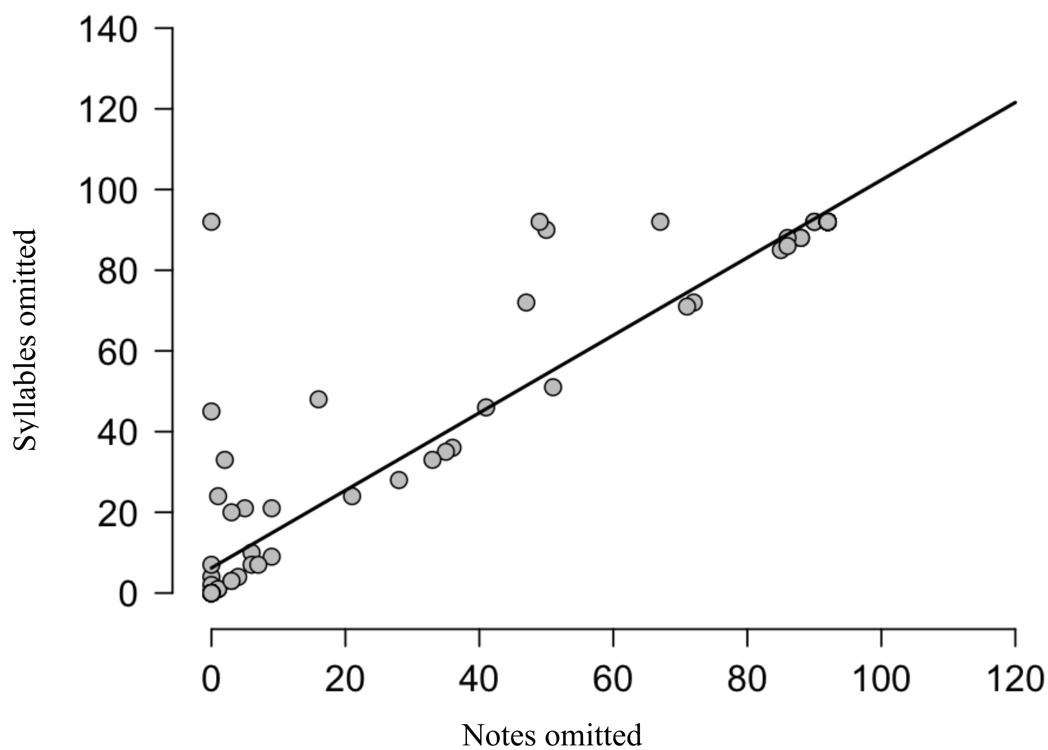

Figure 21. Correlation between syllables and notes omitted at test 2.1.

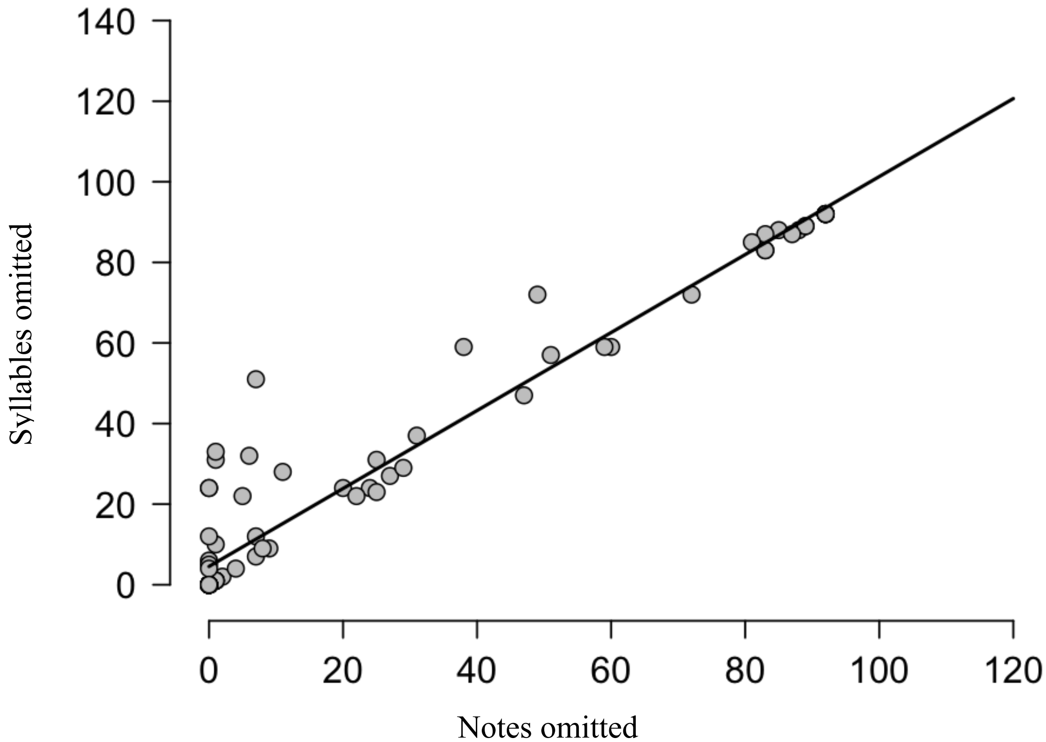

Figure 22. Correlation between syllables and notes omitted at test 2.2.

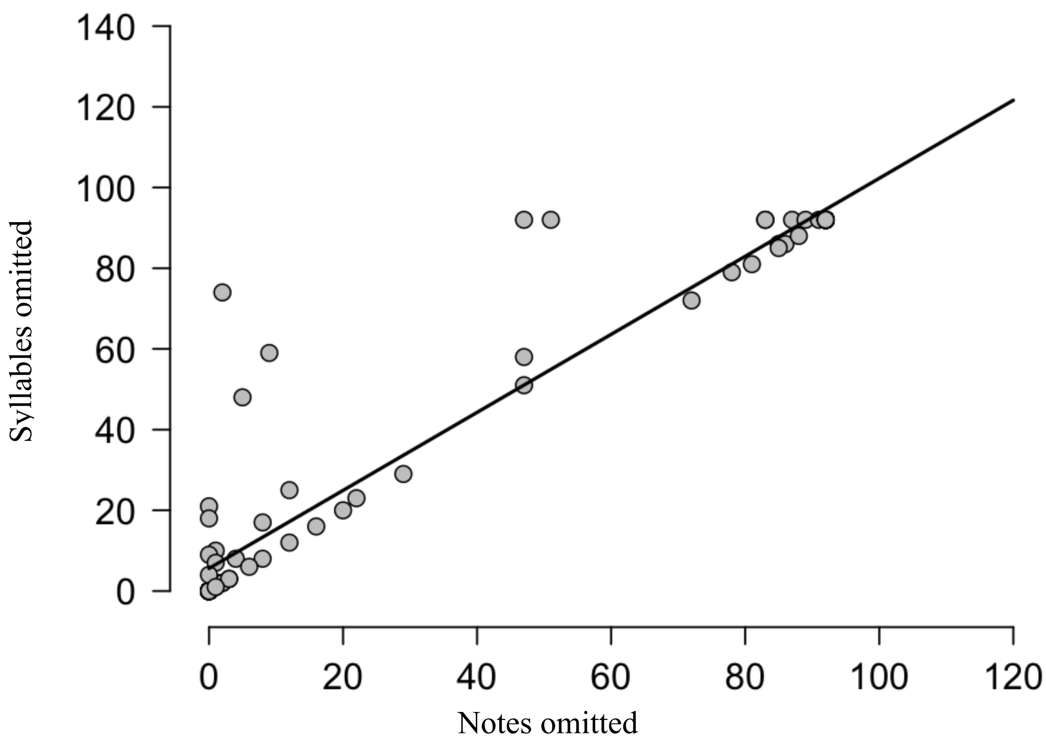

Figure 23. Correlation between syllables and notes omitted at test 3.1.

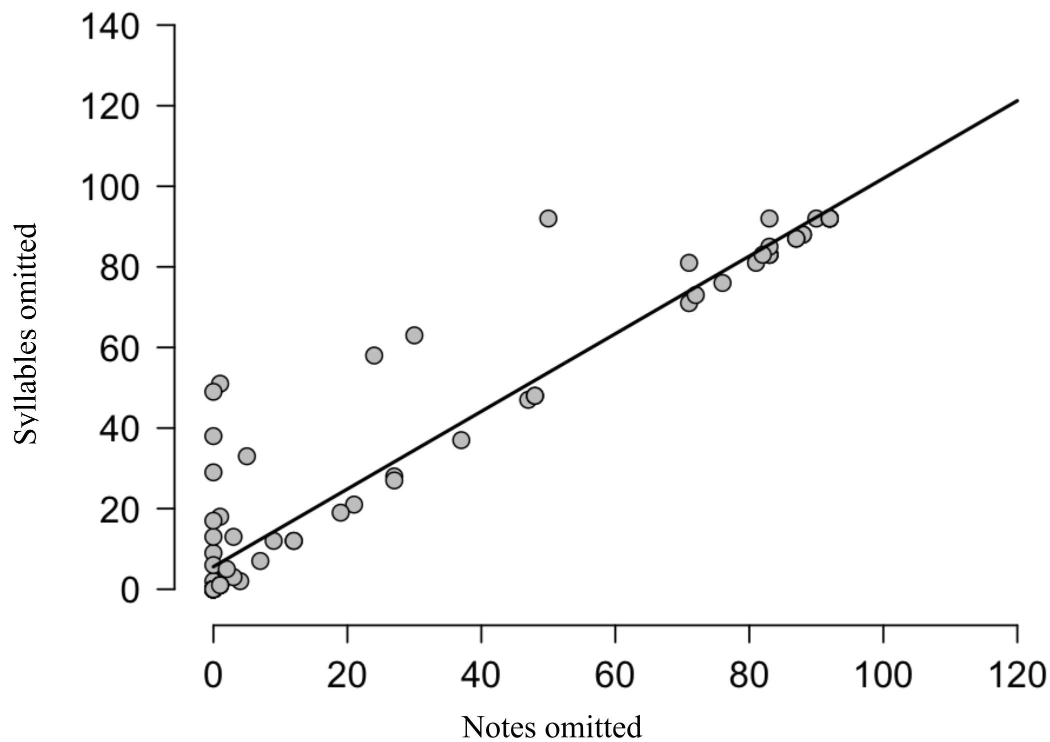

Figure 24. Correlation between syllables and notes omitted at test 3.2

Discussion

5.1 Summary of Findings

The main goal of the experiment was to examine evidence for two hypotheses; that song syllables and notes will show a strong spacing effect, and that there will be no difference in syllable and note retrieval between the two spaced intervals. Two further hypotheses claimed there would be evidence of greater syllable and note forgetting at the two spaced intervals over the massed condition, but evidence of no difference between the two spaced intervals. The hypotheses were derived from research indicating that a song based on traditional materials is remembered through the interaction of multiple constraints relating to poetic materials, musical setting and narrative structure (Wallace & Rubin, 1991). At short (massed) spacing intervals, there will be no need for a reconstructive process from study-phase retrieval (Glenberg, 1979) and hence minimal gain to the associative strength of the links between the syllables of the song and the various constraining features of the material. In the spaced conditions the increased gain in associative strength from study-phase retrieval and the greater probability of overlap in contextual cues between the learning and the testing phase will result in better performance over the massed condition. Once the syllables have been strongly enough associated with the other constraints generated by the ongoing cuing patterns of the song, however, there will be little or no diminishment in retrieval associated with a greater spacing interval. The constraints operate together and offer ongoing cuing of the song which will overtake the usual contextual drift seen at increasing spacing intervals in less-densely associated verbal materials; hence, no Glenberg surface, or inverse U-shaped curve in the spacing effect.

The forgetting curve for sung materials should follow the same logic. Retrieval at short retention intervals will be governed by the associative strength between the notes and the words in conjunction with the constraints of the song materials. This, added to the overlap in contextual cues for the short retention interval, should guarantee excellent recall, once a criterion performance has been achieved. At longer retention intervals, the associative strength between notes and words will be weaker, and the constraining structures of the song materials not sufficiently well-learned to promote effective recall. Moreover, there will be less overlap between contextual cues between the first session and the retrieval. Between the two spaced intervals, however, whatever can be retrieved of the constraining features of the song will tend to diminish the loss of retrieval strength due to increased contextual variability. As a result, there should be evidence of no difference between the two spaced conditions.

Overall, this study found strong evidence for a spacing effect in song. There was evidence that sung text omissions showed a moderate spacing effect at test 3.1 and a very strong spacing effect at test 3.2, with large effect sizes for the difference between spaced and massed conditions at test 3.2 (M vs. S1, $d = 0.873$; M vs. S2, $d = 0.914$). The difference in spacing effects between tests 3.1 and 3.2 can be likely be attributed to differences in cue strength. Session 3.1 tests were cued with the first two notes, but 3.2 tests gave the first two syllables. The syllable cues were more likely to cue episodic traces of the syllables from the learning events, due to transfer-appropriate processing (Lockhart, 2002). Furthermore, at test 3.1, while there was a spacing effect overall between the massed and spaced conditions, and a spacing effect between the massed and one-week interval, there was no spacing effect for the two-day interval compared to the massed. In this case, the poetic constraints of the lyrics may not have been retrieved in response to the note cue. Where structural constraints in the materials are more difficult to retrieve, it is reasonable to expect contextual variability to play more of a role in determining the spacing effect (Glenberg, 1979). The one-week gap will offer a greater variety of contextual cues available for word and note retrieval than the two-day gap; hence a spacing effect for one week but not for two days when tested three weeks later. An examination of means from test 2.1 (Table 9) shows no evidence of difference between the two spaced conditions at the second session initial test, and thus no evidence of differential learning between the two spaced conditions. Because of this, it is less likely that study-phase retrieval can account for differences in the spacing effect at final testing. Once the word cue was given, at test 3.2, spacing effects were strong for both the S1 and S2 condition. This is consistent with the primary hypothesis: that the various constraints of the song allowed greater recall of the song in both spaced conditions compared to massed.

There was no spacing effect for syllable additions and syllable errors, although this may have had to do with floor effects, as there was a very narrow range of data. Once again, examination of mean scores did show a trend toward more additions and errors in the massed compared to the spaced conditions. There was also no evidence of a spacing effect for the number of trials in the third session required to relearn to a criterion of 95% correct syllables. This may have been partially due to the small number of data points available for each condition, as the range of obtained results varied only from 0-6 trials. Examination of mean values for number of trials to criterion did show a tendency toward a greater number of trials to achieve criterion in the third session for the massed condition over the two spaced conditions.

There was a spacing effect for sung notes. Notes omitted overall showed a moderate spacing effect at test 3.1 and an extremely strong spacing effect at test 3.2, with large effect sizes for the difference between spaced and massed conditions at test 3.2 (M vs. S1, $d = 0.957$; M vs. S2, $d = 0.920$). The difference in spacing effects for omitted notes between tests 3.1 and 3.2 may have been due to the effects of omitted syllables. In support of this, Bayesian Pearson correlation showed a strong correlation between notes and words at both initial second session and third session tests. (figures 21-24). There were proportionately more notes than words across all conditions at both final session tests. Participants were asked to produce notes and words for the tests and may have been unwilling to sing notes not accompanied by syllables, except as a placeholder for syllables that were omitted. An examination of individual files showed that where syllables were forgotten, participants sang the notes until they could remember the text. When note data was analyzed separately for quarter notes, dotted quarter and half notes and eighth notes sung, all showed similar spacing patterns with moderate evidence of a spacing effect overall at test 3.1 and strong or extremely strong evidence at test 3.2. In all cases, there were fewer omitted notes at test 3.2, probably because there were fewer omitted words.

There was no evidence of a spacing effect for pitch accuracy measured by cents off-pitch for quarters, dotted quarters and half notes or eighth notes: all notes recalled were sung with approximately equal pitch variation across all conditions. Although there are no studies of distributed learning for tune memory to compare, there is evidence for consistency of long-term tune recall over long retention intervals (Wallace, 1994; Hahn, 2002). Halpern (1989) found that in a population of musicians and non-musicians without absolute pitch, starting pitches for familiar songs were reproduced with considerable consistency across a 48-hour interval.

There was no evidence of a spacing effect for number of notes more than 50 cents off-pitch in either test 3.1 or test 3.2. There also was no evidence of a spacing effect in mean quarter note length or for *SD* of mean quarter note length at test 3.1 or test 3.2. Notes 50 cents or more off-pitch are not just notes out-of-tune: they are also a proxy for wrong notes in the song. That spacing does not affect memory for the notes of a tune at final testing is interesting. Once the notes are well-learned (in this case, subject to two learning sessions), whatever parts of the tune that are reproduced will be retained equally well, however the learning events are distributed.[10] Once again, I could find no studies of distributed learning for sung melody over longer retention intervals that would allow for contextualization of these results. That note length is also independent of spacing effects is not surprising. Note length is a global feature of the rhythmic plan of the piece. Levitin (1996) found that in a group of undergraduates with varying levels of music training 72% were able to come within eight per cent of the remembered tempo of popular songs and were able to reproduce the effect with a high degree of consistency. In this study, quarter note lengths, once understood and associated with the initial cue, were relatively uniform across conditions, differing mainly in the individual tempo set by the singer.

As predicted, there was evidence of no difference in word or tune recall between the two spaced conditions for syllable and note omissions. For syllable omissions, test 3.1 showed no evidence of difference and 3.2 showed moderate evidence of no difference between S1 and S2. As mentioned above, mean scores for S1 at test 3.1 were consistently lower than for S2, without the difference being great enough to constitute evidence of difference. This does suggest that there may indeed be a gap somewhere between 10 minutes and two days that might show a monotonic increase in the spacing function when compared with the ten minute and one-week gaps. There is no such indication of a possible gap effect when the cue is given with words. A further possibility is that a gap of some unspecified length greater than one week could show a decrease in the spacing effect for word retrieval, when cued by notes. A monotonic increase followed by a decline would generate the inverted U-shaped curve (Glenberg surface) often found in spacing studies of less complex verbal materials (Glenberg, 1979). I would not expect

[10] It is important to remember that this is a melody learned from a piano recording and then paired with words, which may have quite a different cognitive profile than a melody learned without words.

to find such a curve generated when both notes and words are given in the cue (test 3.2). In that case, the constraints of the song should be well-enough recovered so that the inverted L-shape found here would apply.

As predicted, there was evidence of no difference in note omission between S1 and S2 at both test 3.1 and at test 3.2. There was evidence of no difference between S1 and S2 for pitch accuracy (cents off-pitch, *SD* of cents off-pitch or *ABS* of cents off-pitch[11] or number of off-pitch quarters). For quarter note length and *SD* of quarter note length, there was mostly evidence of no difference between S1 and S2 at both third session tests. The sole exception was at test 3.1 where there was no evidence of difference for quarter note length. For syllables omitted, notes omitted, quarters sung, dotted quarters and halves sung and eighth notes sung, although there was no evidence of difference, mean scores at test 3.1 for the S1 condition were consistently lower than scores for S2. This presents the interesting possibility that at the two-day ISI words are less closely bound to the two-note melody cue than at the one-week ISI.

Previous research has indicated that words and music in song, although associated in memory, are not so integrated that retrieval of one leads inevitably to retrieval of the other. Listeners who have not been musically trained are more likely to recognize words when sung to originally presented melodies than when sung to altered melodies, indicating association of the original words and tune in memory (Serafine et al., 1986). This association between words and music is robust and extends even to nonsense syllables. Ginsborg (2007) used singers of varying levels of expertise to study memory for a newly-learned unaccompanied song at a ten-minute RI. Two different learning conditions were used, words and tune together and words and tune

---

[11] *SD* of cents off-pitch indicates the degree to which tuning error was distributed across or clustered around certain values; *ABS* indicates the degree of out-of-tune singing.

separate. Errors were analyzed in terms of type, a separate error in either words or tune and a conjoint error where both are incorrect. Nearly twice as many separate errors were made than conjoint errors across all groups. The words could be recalled without the tune, or the tune without the words; it was the learning conditions and degree of expertise of the learner that determined the degree of association between words and notes. In this experiment, it was the inter-study interval interacting with the type of cue and the length of the retention interval that determined how much of the text could be retrieved at final testing.

Forgetting: Fewer syllables were forgotten at M compared to S1 and S2 (Tables 9-11). At both tests 2.1 and 2.2 there was extremely strong evidence for a difference in syllable omission between massed and spaced conditions. There was almost complete recall of the words in the massed condition, with more than half of the words forgotten for the spaced conditions at test 2.1, and one-third to one-half of the words forgotten at test 2.2. There was moderate evidence of no difference between massed and spaced conditions for syllable errors or syllables added at tests 2.1 and 2.2. Mean scores indicated almost no syllable errors or additions in any condition. It is possible that by the second session, the choice of words is already constrained by the song materials to the point where very few syllables are added or sung in error.

There were fewer note omissions at M compared to S1 or S2 (Tables 13 and 14). At both tests 2.1 and 2.2 there was extremely strong evidence that fewer notes were omitted in the massed condition, just as with the syllable omission data. At both tests 2.1 and 2.2 mean scores for note omission showed almost complete recall in the massed condition. At test 2.1 about half of the notes were omitted at S1 and S2; at test 2.2 about one-third of the notes were forgotten at S1 and S2. The same patterns show up in a detailed examination of separate note values (quarters and eighths), with one exception: there was no evidence of spacing for dotted quarter and half notes. This may be due to sparse data. There are only eight instances of dotted quarters and half notes in the song, on the final note of each of the eight lines. Otherwise, there are 76 quarter notes and 12 eighths. With less data, it is more difficult to establish difference.

There was strong evidence for better pitch recall for M compared to S1 and S2 (Tables 19, 20 and 21). At test 2.1, there was very strong evidence of difference for quarter notes off-pitch, and mean scores indicated accuracy within about 5 cents of the ideal pitch for the massed condition, while mean scores were about 3/4ths of a tone flat for S1 and 1/4th of a tone flat for S2. *SD* of cents off-pitch quarters showed extremely strong evidence of difference. *ABS* of cents off-pitch quarters showed strong evidence of difference. Not only was the pitch remembered better at 10 minutes, but there was much less variation in pitch between participants. That pitch recall is more accurate after 10 minutes than after two days or one week is not at all surprising. Long-term accuracy in pitch reproduction has been shown for familiar songs, but not for novel melodies (Halpern, 1989). Very likely, the participants who recalled the song after 10 minutes were able to retrieve the notes, while the others were not. As could be expected, there was no indication that note length was forgotten across conditions.

There was no evidence of difference in syllable omissions between S1 and S2 at test 2.1 or test 2.2. The BF_{10} statistic for both tests was very slightly above the cut-off for moderate evidence of no difference (Table 9). This result is at odds with the usual forgetting curve for word lists (Ebbinghaus 1885/1913). There does not seem to be a forgetting curve in the literature established for poetic materials; however, given the strong overlap between the constraints offered by poetry and that offered by a musical setting of poetry, I would expect the results to be similar. There was moderate evidence of no difference overall for syllables added or syllables in error in tests 2.1 and 2.2. Once again, this speaks to the effect of the overall structural constraints of the material. Once the over-all pattern is understood, participants are not likely to add words that will alter the rhythmic template of the poetry and the music. The occasional error may have been due to the choice of a word that fit the overall rhythmic pattern and offered a similar meaning in context. Number of sung quarters; dotted quarters and half; and eighths all showed exactly the same patterns, moderate evidence of no difference between S1 and S2.

Evidence for a difference in cents off-pitch quarters between 2 days and 1 week was indeterminate (Table 19). Mean values suggest, however, that after two days, pitch on average deviated twice as much as after one week. The same pattern was repeated for the *SD* of cents off-pitch for quarter notes (Table 20) and *ABS* of cents off-pitch for quarter notes. By test 2.4 there was moderate evidence for no difference between S1 and S2 for cents off-pitch quarters, *SD* of cents off-pitch quarters or *ABS* of cents-off pitch quarters. There was no evidence of a difference in number of off-pitch quarter notes (notes greater than 50 cents off-pitch) between conditions at test 2.1. While there were differences in the amount of out-of-tune singing across conditions, it did not affect the number of notes that were more than 50 cents (one quarter note) out of tune. Because more notes were omitted in the spaced conditions, the proportion of notes more than 50 cents out-of-tune was actually higher in the spaced conditions, as indicated by the data for proportion of cents off-pitch quarters. The evidence is that for a single learning session, melody recall after a ten-minute gap is accurate, but equally inaccurate when results two days or one week later are compared. After the second learning session, the tune will be recalled with equal accuracy, regardless of variation in the distribution of learning sessions.

There was no indication of differences in note-length, rests or hesitations between learning conditions at either of the initial second session tests or the final third session tests. Originally, these measures had been included as way of determining hesitation during retrieval.

Note-length may be a poor measure of hesitation, as the final notes before a hesitation would not necessarily show lengthening. Breath length, however, is a measurement which would be expected to capture a hesitation added on to the tempo-related breath between phrases. Considering the few opportunities for hesitation between phrases (there are only seven of these in the stimulus materials) and the number of phrases omitted by the participants at testing, there was not enough data to reach a meaningful determination. In any case, the average breath length may not be a sufficiently sensitive gauge to indicate the effect of hesitation added to the breath. The hesitation measure itself, used where it was possible to separate breath length from hesitation, also generated too few data points to generate a meaningful analysis.

5.2 General discussion

In their review of the spacing literature, Donovan and Radosevich (1999) found significant differences in the spacing effect depending on the type of task. For tasks with high mental requirements, high overall complexity and high physical requirements (a group that included air traffic controller simulation, hand movement memorization and music memorization and performance), the spacing effect was quite weak (d = .07), while for verbal tasks of moderate complexity like free recall, foreign language, and verbal discrimination the effect was moderate (d = 0.42), and for simple mechanical tasks the effect was strong (d = 0.97). For verbal tasks, increasing the study interval between distributed practice events improved recall at final tests until a certain optimal point. Beyond that, results followed an inverse U function, also known as the Glenberg surface; the effect of distributed learning diminished, and smaller effect sizes were found. For the task group including music performance increasing the time between study events had a different effect; the larger the lag between study events, the worse the performance. Overall, only the task groups containing tasks of high and moderate complexity (containing both the verbal tasks and musical performance) could be evaluated for retention intervals greater than one day. For these two groups, longer retention intervals led to poorer performance.

There are a number of problems with this from a musician's standpoint. To begin with, the legitimacy of grouping music performance with air traffic controller simulation and hand movement memorization is doubtful. The cognitive profile of these particular activities may be quite different. A close examination of the reference materials included in the Donovan and Radosevich indicated that only a single music study was included in the meta-analysis, a doctoral dissertation by Yoas (1982) on distributed learning for instrumentalists. 'Music memorization and performance' is itself a generalization that must be unpacked if it is to have any specific meaning. Memory for song, with its linguistic component and memory for the various textures of instrumental music may all show different cognitive profiles and should all be studied individually. Performance memory for instruments adapted to polyphony, those that are primarily melodic, and those primarily rhythmic may exhibit different characteristics. As I have argued above, the various constraints of sung materials also have an important effect on memory for song and may allow for long-term verbatim retrieval of complex verbal patterns (Rubin, 1995).

Cepeda et al. 2006 conducted a meta-analysis of the literature on distributed learning for verbal materials, looking at the interaction of retention interval, inter-study interval and task. They found that as the retention interval increased, the greatest increases in test performance were found at longer ISI differences. Overall, they also found "a non-monotonic effect of absolute ISI upon memory performance at a given retention interval", in other words, the classic inverted U-shape curve that drops off after an optimal point is reached for any retention interval

(Glenberg, 1979). This was confirmed by a large internet-based study of fact learning at retention intervals of up to 350 days (Cepeda et al., 2008)

This dissertation research looked at song that follows the traditional ballad structure and found that increasing the study interval between distributed practice events produced an inverted L-shape, where memory improves between massed (10-minute lag) and spaced sessions but plateaus between sessions presented at two days and at one week. What is not known, however, is what would happen to memory for this particular type of song at increasingly spaced intervals. It is indeed possible for any given RI, that recall would drop off if the lag between learning events were sufficiently large.

Mean scores at final testing indicate that where the notes and words are cued by notes only, the increase in mean scores for words recalled is monotonic, although not showing a statistical difference between the two spaced conditions. At test 3.2, however, where the cue was by notes and words, mean scores show a clear plateau for S1 and S2. Any prediction of the spacing effect at increasing lags would be guess-work; but I am inclined to think that there would eventually be a drop-off in words recalled at test 3.1 due to increasingly weak retrieval of words associated with the tune, and loss of recall for the notes of the tune. At test 3.2 memory for the constraints triggered by the initial words could well be enough to continue to enhance retrieval of syllables and notes, albeit at a lower plateau.

In this study, results for notes and results for syllables were linked, which may indicate that participants were indeed obeying instructions, and consistently attempting to produce words, even when cued by notes. The consistently larger scores for syllables as compared to notes omitted in all second and third session memory tests are consistent with the notes being sung without words as a placeholder when the words were forgotten. In the future, it will be important to look at tune memory alone and at memory for the lyrics without the tune in order to dissociate the effects of words from music in song recall.

The spacing effect in this study turned up as a difference in scores for words retrieved between the ten-minute and two-day/one-week ISIs. Previous research on verbal materials has demonstrated the relationship between inter-study interval and retention interval. In particular, optimal ISIs are shorter for short RIs, and longer for long RIs. Moreover, for a range of RIs of up to 28 days, a gap of 24 hours between learning sessions generally leads to better recall than shorter ISIs. (Cepeda et al., 2006, p. 362). This brings up the interesting possibility that a period of sleep may be important in amplifying the spacing effect. In a study of long-term memory for Swahili-English word pairs, Bell et al. (2014) compared inter-study intervals that were massed, or spaced at 12 hours without a sleep interval, at 12 hours with a sleep interval and at 24 hours (Bell, Kawadri, Simone and Wiseheart, 2014). They found that the 12-hour gap with sleep and the 24-hour gap both produced significantly better scores at a ten-day retention interval than the massed or 12 hour no-sleep condition. The authors theorized that the critical factor for improved recall was not the consolidation of memory during sleep, but the increased difficulty of retrieval of the material after the sleep interval (i.e. study-phase retrieval). Similarly, in this study, the spacing effects observed for words recalled from the note and word cue at final testing (test 3.2) were due in part to more difficult second session recall for the words in the spaced conditions, i.e. study-phase retrieval. Whether the period of sleep between learning events is the crucial determinant is a subject for future research.

5.3 Importance and Relevance

There are many direct applications for a study of song spacing. The optimal distribution of song learning practice among a normal population is relevant as a comparative baseline for musical mnemonic training, (MMT), a therapeutic approach to memory (Murakami, 2017). Foreign language training can also benefit from song spacing research. A recent study of singing as an aid to foreign language learning in a population of school age children (Good, Russo & Sullivan, 2014) used a learning schedule of six sessions distributed over two weeks. This was a choice based on the hands-on experience of teaching English as a second language. There is currently no empirical research (other than this study) to support their decision. Singers preparing for performance currently have no idea what practice schedule they should use to better guarantee fluent memory retrieval during performance. This research offers the possibility of concrete suggestions to guide them.

5.4 Limitations

Because there is no exact parallel for this study in the literature, certain parameters of the study were of necessity determined by assumption. The spacing intervals of ten minutes, two days and one week were chosen based on Cepeda, Vul, Rohrer, Wixted & Pashler, 2008, with the expectation that they represented a distribution that would capture an effect. The retention interval of three weeks was chosen to accommodate two training sessions and a testing session for 90 participants over the course of a single academic semester of 12 weeks. The two verse, eight-line structure was chosen to be learnable to criterion in a single session of forty minutes, without being so easy to recall that ceiling effects would predominate. The text was an attempt to write two verses that would be of interest to undergraduates. The tune was greatly simplified over the original inspiration, a traditional Nova Scotia sea shanty, although the rhyming structure and accent pattern of the original text was maintained. The two note or two syllable and note cues were chosen to allow the maximum syllable and note evaluation possible; a six note or note-and-syllable cue might have produced different results. Although there was no indication statistically of difference between conditions for melody or rhythm of the stimulus materials in this study, in the future, stimulus materials exactly matched for pitch and rhythm should be used. Similarly, although there was statistically no indication of difference between conditions for English speakers, there were differences in mean numbers between conditions. It would be better to distribute non-native English speakers equally between conditions. Finally, although I was blind to condition while coding the data, I was not blind to condition during the coding phase. A script was followed and every effort was made to maintain uniformity across conditions. Differences in spoken vs. sung coaching styles in the first session were tracked and no difference was found across conditions. Going forward, every effort should be made to elicit only sung responses.

5.5 Future research

To my knowledge, there have been no studies of distributed learning for long term memory of lyric text (poetry) not set to music. It would be extremely interesting to compare the results of this study with a study of the words without the tune and the tune without the words over similar gaps and retention intervals. That way, the relative contribution of melody to word memory for a song could be assessed. Multiple learning episodes should also be studied, since they would correspond better with real-life classroom learning situations. An internet-based study of song learning could be designed that would allow for more participation and the study of a variety of inter-study intervals and retention intervals. Classroom-based song learning could be studied, perhaps over multiple learning trials equally spaced (fixed intervals) or distributed more towards the beginning of the study period (expanding intervals) or toward the end of the study period (contracting intervals) (Küpper-Tetzel, Kapler & Wiseheart, 2014). It would be interesting to study expert singers in preparation for performance, to examine the role of descriptive, structural and contextual cues in performance memory. Studies of traditional methods of committing sacred text to memory could also amplify the work of Rubin and Wallace on traditional ballads (Rubin, 1995; Rubin, 2006; Wallace 1994; Wallace & Rubin, 1988a; Wallace & Rubin, 1988b; Wallace & Rubin, 1991).

5.6 Conclusions

This study adds to a growing body of knowledge of the cognitive profile of song. The experiment conducted here established that when learning sessions are distributed over several days, verbatim memory for sung text can be robust over an interval of three weeks, and possibly, much longer. In contrast, massed learning for sung (up to the ten-minute gap used in this study) was ineffective as a way of committing song to memory, at least for the three-week retention interval we examined. The exact spacing interval seems to matter less than the necessity of a sufficient interval to promote long-term retention. The parameters of that interval are not yet exactly known, but clearly, ten minutes between learning sessions are not enough and two days are enough, at least for a retention interval of three weeks. Considering the interest in song as a rehabilitative tool, as a means of teaching vocabulary in second language instruction, and as a vehicle for retrieving autobiographical memory, we need to know more about what contributes to memory for song, and what inhibits its retention.

References

Aldridge, D. (Ed.) (2005). *Music therapy and neurological rehabilitation: Performing health*. London, UK: J. Kingsley Publishers.

Atchley, R. M., & Hare, M. L. (2013). Memory for poetry: More than meaning? *International Journal of Cognitive Linguistics*, *4*(1), 35-50. Retrieved from https://www-ncbi-nlm-nih-gov./pmc/articles/PMC4577018/

Atkinson, R. C., & Shiffrin, R. M. (1968). Human memory: A proposed system and its control processes. In K. Spence & J.T. Spence (Eds.) *Psychology of learning and motivation, Vol. 2* (pp. 89-195). Cambridge, MA: Academic Press.

Baddeley, A. (2012). Working memory: theories, models, and controversies. *Annual Review of Psychology*, *63*, 1-29. https://doi.org/10.1146/annurev-psych-120710-100422

Baddeley, A. D. (2004). The psychology of memory. In A. D. Baddeley, M. Kopelman & B.A. Wilson (Eds.), *The essential handbook of memory disorders for clinicians* (pp. 1-13). Chichester, UK: John Wiley & Sons.

Baddeley, A. (2000). The episodic buffer: a new component of working memory? *Trends in Cognitive Sciences*, *4*(11), 417-423. https://doi.org/10.1016/S1364-6613(00)01538-2

Baddeley, A. D., & Hitch, G. J. (1974). Working memory. *The psychology of learning and motivation, 8*, 47-89. https://doi.org//10.1016/S0079-7421(08)60452-1

Bartlett, F. C. (1932). *Remembering: An experimental and social study*. Cambridge, UK: Cambridge University Press.

Bell, M. C., Kawadri, N., Simone, P. M., & Wiseheart, M. (2014). Long-term memory, sleep, and the spacing effect. Memory, 22(3), 276-283. https://doi.org/10.1080/09658211.2013.778294

Bellezza, F. S. (1996). Mnemonic methods to enhance storage and retrieval. In E.I. Bjork & R.A. Bjork (Eds.), *Memory*, (pp. 345-380). Cambridge, MA: Academic Press.

Berz, W. L. (1995). Working memory in music: A theoretical model. *Music Perception: An Interdisciplinary Journal*, *12*(3), 353-364. https://doi.org/10.2307/40286188

Bjork, R. A., & Bjork, E. L. (1992). A new theory of disuse and an old theory of stimulus fluctuation. In A. F. Healy, S. M. Kosslyn, & R. M. Shiffrin (Eds.), *From learning processes to cognitive processes: Essays in honor of William K. Estes: Vol. 2* (pp. 35-67): Cambridge, MA: Psychology Press.

Bjork, R. A., Bjork, E. L., & Caughey, J. B. (2007). Retrieval as a self-limiting process: Part II. In J. S. Nairne (Ed.), *The foundations of remembering: Essays in honor of Henry L. Roediger, III* (pp. 19-37). Cambridge, MA: Psychology Press.

Boas, F. (1898/2006). *The mythology of the Bella Coola Indians*. Whitefish, MT: Kessinger Publishing.

Boltz, M. (1991). Some structural determinants of melody recall. *Memory & Cognition*, 19(3), 239-251. https://doi.org/10.3758/BF03211148

Bower, G. H. (2000). A Brief History of Memory Research. In E. Tulving & F. I. M. Craik (Eds.), *The Oxford handbook of memory* (pp. 3-33). Oxford, UK: Oxford University Press.

Brown, S. C., & Craik, F. I. (2000). Encoding and retrieval of information. In E. Tulving & F. I. M. Craik. (Eds.), *The Oxford handbook of memory* (pp. 93-107). Oxford, UK: Oxford University Press.

Byron, R. (1995). The enthnomusicology of John Blacking. In R. Byron (Ed.), *Music, culture, and experience: Selected papers of John Blacking* (pp. 1-31). Chicago, IL: University of Chicago Press.

Cady, E. T., Harris, R. J., & Knappenberger, J. B. (2008). Using music to cue autobiographical memories of different lifetime periods. *Psychology of Music*, 36(2), 157-178. https://doi.org/10.1177/0305735607085010

Calvert, S. L., & Billingsley, R. L. (1998). Young children's recitation and comprehension of information presented by songs. *Journal of Applied Developmental Psychology*, *19*(1), 97-108. https://doi.org/10.1016/S0193-3973(99)80030-6

Calvert, S. L., & Tart, M. (1993). Song versus verbal forms for very-long-term, long-term, and short-term verbatim recall. *Journal of Applied Developmental Psychology*, *14*(2), 245-260. https://doi.org/10.1016/0193-3973(93)90035-T

Carpenter, S. K., Cepeda, N. J., Rohrer, D., Kang, S. H., & Pashler, H. (2012). Using spacing to enhance diverse forms of learning: Review of recent research and implications for instruction. *Educational Psychology Review*, 24(3), 369-378. https://doi.org/10.1007/s10648-012-9205-z

Cash, C. D. (2009). Effects of early and late rest intervals on performance and overnight consolidation of a keyboard sequence. *Journal of Research in Music Education*, *57*(3), 252-266. https://doi.org/10.1177/0022429409343470

Cepeda, N. J., Pashler, H., Vul, E., Wixted, J. T., & Rohrer, D. (2006). Distributed practice in verbal recall tasks: A review and quantitative synthesis. *Psychological Bulletin, 132*(3), 354-380. https://doi.org/10.1037/0033-2909.132.3.354

Cepeda, N. J., Vul, E., Rohrer, D., Wixted, J. T., & Pashler, H. (2008). Spacing effects in learning: A temporal ridgeline of optimal retention. *Psychological Science, 19*(11), 1095-1102. https://doi.org/10.1111/j.1467-9280.2008.02209.x

Chaffin, R., & Imreh, G. (2002). Practicing perfection: Piano performance as expert memory. *Psychological Science, 13*(4), 342-349. https://doi.org/10.1111/j.0956-7976.2002.00462.x

Chaffin, R., Logan, T. R., & Begosh, K. T. (2008). Performing from memory. In S. Hallam, I. Cross, & M. Thaut (Eds.), *Oxford handbook of music psychology*. Oxford, UK: Oxford University Press.

Chazin, S., & Neuschatz, J. S. (1990). Using a mnemonic to aid in the recall of unfamiliar information. *Perceptual and Motor Skills*, *71*(3 suppl.), 1067-1071. https://doi.org/10.2466/pms.1990.71.3f.1067

Cirigliano, M. M. (2013). Musical mnemonics in health science: A first look. *Medical Teaching, 35*(3), e1020-e1026. https://doi.org/10.3109/0142159X.2012.733042

Cofer, C. N. (1941). A comparison of logical and verbatim learning of prose passages of different lengths. *American Journal of Psychology*, 54(1), 1-20. https://doi.org/10.2307/1417789

Conway, M. A. (2001). Sensory–perceptual episodic memory and its context: Autobiographical memory. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences, 356* (1413), 1375-1384. https://doi.org/10.1098/rstb.2001.0940

Conway, M. A., & Pleydell-Pearce, C. W. (2000). The construction of autobiographical memories in the self-memory system. *Psychological Review, 107*(2), 261-288. https://doi.org//10.1037/0033-295X.107.2.261

Craik, F. I. M., & Lockhart, R. S. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior, 11*(6), 671-684. https://doi.org/10.1016/S0022-5371(72)80001-X

Creighton, H., & Senior, D. H. (1950). *Traditional Songs from Nova Scotia*. Toronto, ON: Ryerson Press.

Crowder, R. G. (1976/2014). *Principles of Learning and Memory: Classic Edition*. Cambridge, MA: Psychology Press.

Cuddy, L. L., Cohen, A. J., & Mewhort, D. J. K. (1981). Perception of structure in short melodic sequences. *Journal of Experimental Psychology: Human Perception and Performance*, *7*(4), 869-883. https://doi.org/10.1037/0096-1523.7.4.869

Delaney, P. F., Verkoeijen, P. P., & Spirgel, A. (2010). Spacing and testing effects: A deeply critical, lengthy, and at times discursive review of the literature. In B. H. Ross (Ed.), *The psychology of learning and motivation: Advances in research and theory: Vol. 53* (pp. 63-147). New York, NY: Academic Press.

DeNora, T. (2000). *Music in everyday life*. Cambridge, UK: Cambridge University Press.

Dowling, W. J. (1978). Scale and contour: Two components of a theory of memory for melodies. *Psychological Review*, 85(4), 341-354. https://doi.org/10.1037/0033-295X.85.4.341

Ebbinghaus, H. (1885/1913). *Memory: A contribution to experimental psychology* (H. A. Ruger, C. E. Bussenius, & E. R. Hilgard, Trans.). Teachers College Press. Retrieved from https://books-scholarsportal-info.ezproxy.library.yorku.ca/uri/ebooks/ebooks0/apa/2010-03-04/1/200416275

Engle, R.W. (2001). What is working memory capacity? In H.L. Roediger, J.S. Nairne, I. Neath, & A.M. Suprenant (Eds.), *The nature of remembering: Essays in honor of Robert G. Crowder* (pp. 297–314). Washington, DC: American Psychological Association Press.

Engle, R.W., & Oransky, N. (1999). Multi-store versus dynamic models of temporary storage in memory. In R.J. Sternberg (Ed.), *The nature of cognition* (pp. 514–555). Cambridge, MA: MIT Press.

Finke, C., Esfahani, N. E., & Ploner, C. J. (2012). Preservation of musical memory in an amnesic professional cellist. *Current Biology*, *22*(15), R591-R592. https://doi.org/10.1016/j.cub.2012.05.041

Ginsborg, J. (2002). Classical singers learning and memorising a new song: An observational study. *Psychology of Music, 30*(1), 58-101. https://doi.org/10.1177/0305735602301007

Ginsborg, J. (2004). Strategies for memorizing music. In A. Williamon (Ed.), *Musical excellence: Strategies and techniques to enhance performance* (pp. 123-141). New York, NY: Oxford University Press.

Ginsborg, J., & Chaffin, R. (2011). Performance cues in singing: Evidence from practice and recall. In *Music and the mind: Investigating the functions and processes of music (a book in honour of John Sloboda)* (pp. 339-360). Oxford, UK: Oxford University Press.

Ginsborg, J., & Sloboda, J. A. (2007). Singers' recall for the words and melody of a new, unaccompanied song. *Psychology of Music, 35*(3), 421-440. https://doi.org/10.1177/0305735607072654

Glenberg, A. M. (1979). Component-levels theory of the effects of spacing of repetitions on recall and recognition. *Memory and Cognition, 7*(2), 95-112. https://doi.org/10.3758/BF03197590

Glenberg, A.M. (1997). What Memory is for. *Behavioral and Brain Sciences* 20, 1-55. Retrieved from https://www-cambridge-org.ezproxy.library.yorku.ca/core/services/aop-cambridge-core/content/view/46E0F728DAA67D54AA44D2C4634D9556/S0140525X97000010a.pdf/what-memory-is-for.pdf

Good, A. J., Russo, F. A., & Sullivan, J. (2014). The efficacy of singing in foreign-language learning. *Psychology of Music, 43*(5), 627–640. https://doi.org/10.1177/0305735614528833

Hahn, K. (2002). Absolute memory for melody: A study with choir singers. In C. Stevens, D. Burnham, G. McPherson, E. Schubert & J. Renwick (Eds.), *Proceedings of the 7th international conference on music perception & cognition*, 514–517. Adelaide, AU: Causal Productions.

Hallam, S. (2010). The power of music: Its impact on the intellectual, social and personal development of children and young people. *International Journal of Music Education, 28*(3), 269-289. https://doi.org/10.1177/0255761410370658

Halpern, A. R. (1989). Memory for the absolute pitch of familiar songs. *Memory and Cognition*, *17*(5), 572-581. https://doi.org/10.3758/BF03197080.

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

Kim, A. S. N., Wong-Kee-You, A. M. B., Wiseheart, M., & Rosenbaum, R. S. (2019). The spacing effect stands up to big data. *Behavior Research Methods*, 1-13. https://doi.org/10.3758/s13428-018-1184-7

Kintsch, W., & van Dijk, T. (1978). Cognitive psychology and discourse: Recalling and summarizing stories. In Dressler, W. U. (Ed.), *Current trends in text linguistics (*pp. 61-80). Berlin, BRD: De Gruyter.

Küpper-Tetzel, C. E., Kapler, I. V., & Wiseheart, M. (2014). Contracting, equal, and expanding learning schedules: the optimal distribution of learning sessions depends on retention interval. *Memory Cognition, 42*(5), 729-741. https://doi.org/10.3758/s13421-014-0394-1

Küpper-Tetzel, C. E., & Erdfelder, E. (2012). Encoding, maintenance, and retrieval processes in the lag effect: a multinomial processing tree analysis. *Memory, 20*(1), 37-47. https://doi.org/10.1080/09658211.2011.631550

Law, L. N., & Zentner, M. (2012). Assessing musical abilities objectively: Construction and validation of the Profile of Music Perception Skills. *PloS one*, *7*(12). Retrieved from https://doi.org/10.1371/journal.pone.0052508.

Lea, R. B., Rapp, D. N., Elfenbein, A., Mitchel, A. D., & Romine, R. S. (2008). Sweet silent thought: Alliteration and resonance in poetry comprehension. *Psychological Science*, *19*(7), 709-716. https://doi.org/10.1111/j.1467-9280.2008.02146.x

Lee, T. D., & Genovese, E. D. (1988). Distribution of practice in motor skill acquisition: Learning and performance effects reconsidered. *Research Quarterly for Exercise and Sport, 59*(4), 277-287. https://doi.org/10.1080/02701367.1988.10609373

Levitin, D. J., & Cook, P. R. (1996). Memory for musical tempo: Additional evidence that auditory memory is absolute. *Perception & Psychophysics*, *58*(6), 927-935. https://doi.org/10.3758/BF03205494

Live Nation Entertainment. https://www.setlist.fm/stats/average-setlist/bruce-springsteen-2bd6dcce.html?year=2018. Retrieved 10:54 a.m. Jan. 12, 2019.

Lockhart, R.S. (2002). Levels of processing, transfer-appropriate processing, and the concept of robust encoding, *Memory*, 10(5-6), 397-403. https://doi.org/10.1080/09658210244000225

Ludke, K. M. (2018). Singing and arts activities in support of foreign language learning: an exploratory study. *Innovation in Language Learning and Teaching, 12*(4), 371-386. https://doi.org/10.1080/17501229.2016.1253700

Martin, R. C., Wogalter, M. S., & Forlano, J. G. (1988). Reading comprehension in the presence of unattended speech and music. *Journal of Memory and Language*, *27*(4), 382-398. https://doi.org/10.1016/0749-596X(88)90063-0

McElhinney, M., & Annett, J. M. (1996). Pattern of efficacy of a musical mnemonic on recall of familiar words over several presentations. *Perceptual and Motor Skills*, *82*(2), 395-400. https://doi.org/10.2466/pms.1996.82.2.395

Meyer, B. J., & McConkie, G. W. (1973). What is recalled after hearing a passage? *Journal of Educational Psychology*, 65(1), 109-117. http://dx.doi.org/10.1037/h0034762

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review, 63*(2), 81-97. http://dx.doi.org/10.1037/h0043158

Mishra, J., & Backlin, W. M. (2007). The effects of altering environmental and instrumental context on the performance of memorized music. *Psychology of Music*, *35*(3), 453-472. https://doi.org/10.1177/0305735607077838

Miyake, A., & Shah, P. (Eds.). (1999). *Models of working memory: Mechanisms of active maintenance and executive control*. Cambridge, UK: Cambridge University Press.

Moss, V. D. (1995). *The efficacy of massed versus distributed practice as a function of desired learning outcomes and grade level of the student* (Doctoral dissertation). Retrieved from https://digitalcommons.usu.edu/etd/3552

Murakami, B. (2017). *Music as a mnemonic device for verbal recall in healthy older adults* (Master's thesis). Retrieved from http://scholarlyrepository.miami.edu/oa_theses/666

Neubäcker, P. (2011). *Sound-object oriented analysis and note-object oriented processing of polyphonic sound recordings*. U.S. Patent 8,022,286. Retrieved from https://patents-google-com.ezproxy.library.yorku.ca/patent/US8022286B2/en

Noice, T., & Noice, H. (1997). *The nature of expertise in professional acting: A cognitive view*. London, UK: Psychology Press.

Noice, T., & Noice, H. (2002). Very long-term recall and recognition of well-learned material. *Applied Cognitive Psychology, 16*(3), 259-272. https://doi.org/10.1002/acp.785

Noice, H., Noice, T., & Kennedy, C. (2000). Effects of enactment by professional actors at encoding and retrieval. *Memory, 8*(6), 353-363. https://doi.org/10.1080/09658210050156813

Oura, Y., & Hatano, G. (1988). Memory for melodies among subjects differing in age and experience in music. *Psychology of Music*, *16*(2), 91-109. https://doi.org/10.1177/0305735688162001

Pashler, H., Cepeda, N., Lindsey, R. V., Vul, E., & Mozer, M. C. (2009). Predicting the optimal spacing of study: A multiscale context model of memory. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, & A. Culotta (Eds.), *Advances in neural information processing systems: Vol. 22* (pp. 1321-1329). La Jolla, CA: Neural Information Processing Systems Foundation.

Patel, A. D. (2010). *Music, language, and the brain*. Oxford, UK: Oxford University Press.

Paul, I. H. (1959). *Studies in remembering: The reproduction of connected and extended verbal material*. New York, NY: International Universities Press.

Povel, D. J., & Jansen, E. (2001). Perceptual mechanisms in music processing. *Music Perception: An Interdisciplinary Journal*, *19*(2), 169-197. https://doi.org/10.1525/mp.2001.19.2.169-197

Purnell-Webb, P., & Speelman, C. P. (2008). Effects of music on memory for text. *Perceptual and Motor Skills*, *106*(3), 927-957. https://doi.org/10.2466/pms.106.3.927-957

Racette, A., & Peretz, I. (2007). Learning lyrics: To sing or not to sing? *Memory and Cognition, 35*(2), 242-253. https://doi.org/10.3758/BF03193445

Rainey, D. W., & Larsen, J. D. (2002). The effect of familiar melodies on initial learning and long-term memory for unconnected text. *Music Perception: An Interdisciplinary Journal*, *20*(2), 173-186. https://doi.org/10.1525/mp.2002.20.2.173

Rapp, D. N., & Samuel, A. G. (2002). A reason to rhyme: Phonological and semantic influences on lexical access. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*(3), 564-571. https://doi.org/10.1037/0278-7393.28.3.564

Rawson, K. A., & Kintsch, W. (2005). Rereading effects depend on time of test. *Journal of Educational Psychology*, 97(1), 70-80. https://doi.org/10.1037/0022-0663.97.1.70

Rubin, D. C. (1995). *Memory in oral traditions: The cognitive psychology of epic, ballads, and counting-out rhymes*. New York, NY: Oxford University Press.

Rubin, D. C. (2006). The basic-systems model of episodic memory. *Perspectives on Psychological Science, 1*(4), 277-311. https://doi.org/10.1111/j.1745-6916.2006.00017.x

Rubin, D. C., Wallace, W. T., & Houston, B. C. (1993). The beginnings of expertise for ballads. *Cognitive Science*, 17(3), 435-462. https://doi.org/10.1207/s15516709cog1703_4

Rubin, D. C., & Wallace, W. T. (1989). Rhyme and reason: Analyses of dual retrieval cues. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *15*(4), 698-709. https://doi.org/10.1037/0278-7393.15.4.698

Rubin-Rabson, G. (1940). Studies in the psychology of memorizing piano music: II. A comparison of massed and distributed practice. *Journal of Educational Psychology, 31*(4), 270.-284. https://doi.org/10.1037/h0061174

Sachs, J. S. (1967). Recognition memory for syntactic and semantic aspects of connected discourse. *Perception & Psychophysics*, *2*(9), 437-442. https://doi.org/10.3758/BF03208784

Schmuckler, M. A. (2016). Tonality and contour in melodic processing. In S. Hallam, I. Cross & M. Thaut (Eds.), *The Oxford handbook of music psychology* (pp. 143-166). New York, NY: Oxford University Press.

Schön, D., Boyer, M., Moreno, S., Besson, M., Peretz, I., & Kolinsky, R. (2008). Songs as an aid for language acquisition. *Cognition*, *106*(2), 975-983. https://doi.org/10.1016/j.cognition.2007.03.005

Segalowitz, N., Cohen, P., Chan, A., & Prieur, T. (2001). Musical recall memory: Contributions of elaboration and depth of processing. *Psychology of Music, 29*(2), 139-148. https://doi.org/10.1177/0305735601292004

Serafine, M. L., Davidson, J., Crowder, R. G., & Repp, B. H. (1986). On the nature of melody-text integration in memory for songs. *Journal of Memory and Language*, *25*(2), 123-135. https://doi.org/10.1016/0749-596X(86)90025-2

Simmons, A. L. (2012). Distributed practice and procedural memory consolidation in musicians' skill learning. *Journal of Research in Music Education*, *59*(4), 357-368. https://doi.org/10.1177/0022429411424798

Simmons, A. L., & Duke, R. A. (2006). Effects of sleep on performance of a keyboard melody. *Journal of Research in Music Education*, *54*(3), 257-269 https://doi.org/10.1177/002242940605400308

Sloboda, J. A. (1999). Everyday uses of music listening: A preliminary study. In S.W. Yi (Ed.), *Music, mind and science*, (pp. 354-369). Seoul, SK: Seoul University Press.

Sloboda, J. A., O'Neill, S. A., & Ivaldi, A. (2001). Functions of music in everyday life: An exploratory study using the Experience Sampling Method. *Musicae Scientiae, 5*(1), 9-32. https://doi.org/10.1177/102986490100500102

Sgroi, C. (2005). Song cycle translations. Retrieved from http://www.gopera.com/lieder/translations/schubert.html

Snyder, B. (2008). Music and memory. In S. Hallam, I. Cross, & M. Thaut (Eds.), *Oxford handbook of music psychology* (pp. 107-117). Oxford, UK: Oxford University Press.

Stambaugh, L. A. (2011). When repetition isn't the best practice strategy: Effects of blocked and random practice schedules. *Journal of Research in Music Education*, *58*(4), 368-383. https://doi.org/10.1177/0022429410385945

Stambaugh, L. A., & Demorest, S. M. (2010). Effects of practice schedule on wind instrument performance: a preliminary application of a motor learning principle. *Update: Applications of Research in Music Education*, *28*(2), 20-28. https://doi.org/10.1177/8755123310361768

Ten Berge, T., & Van Hezewijk, R. (1999). Procedural and declarative knowledge: An evolutionary perspective. *Theory & Psychology*, *9*(5), 605-624. https://doi.org/10.1177/0959354399095002

Tillmann, B., & Dowling, W. J. (2007). Memory decreases for prose, but not for poetry. *Memory and Cognition, 35*(4), 628-639. https://doi.org/10.3758/BF03193301

Tulving, E. (2002). Episodic memory: From mind to brain. *Annual Review of Psychology, 53*(1), 1-25. https://doi.org/10.1146/annurev.psych.53.100901.135114

Tulving, E., & Thomson, D. M. (1973). Encoding specificity and retrieval processes in episodic memory. *Psychological Review, 80*(5), 352. https://doi.org/10.1037/h0020071

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

Wallace, W.T., & Rubin, D.C. (1988a). Memory of a ballad singer. In M.M. Gruneberg, P.E. Morris & R.N. Sykes (Eds.), *Practical aspects of memory: Current research and issues, Vol. 1. Memory in everyday life* (pp. 257-262). Oxford, UK: John Wiley & Sons.

Wallace, W. T., & Rubin, D. C. (1988b). "The wreck of the old 97": A real event remembered in song. In U. Neisser & E. Winograd (Eds.), *Remembering reconsidered: Ecological and traditional approaches to the study of memory* (pp. 283-310). New York, NY: Cambridge University Press.

Wallace, W. T., & Rubin, D. C. (1991). Characteristics and constraints in ballads and their effects on memory. *Discourse Processes, 14*(2), 181-202. https://doi.org/10.1080/01638539109544781

Warrington, E. K., & Weiskrantz, L. (1968). New method of testing long-term retention with special reference to amnesic patients. *Nature*, 217(5132), 972-974. https://doi.org/10.1038/217972a0

Weston, C. (2018). *Putting the distributed practice effect into context.* (Unpublished doctoral dissertation). York University, Toronto.

White, N. I. (Ed.). (1952-1962). *The Frank C. Brown collection of North Carolina folklore (Vols. 2 -5)*. Durham, NC: Duke University Press.

Wiseheart, M., D'Souza, A. A., & Chae, J. (2017). Lack of spacing effects during piano learning. *PloS One*, *12*(8): e0182986. https://doi.org/10.1371/journal.pone.0182986

Wiseheart, M., Küpper-Tetzel, C. E., Weston, T., Kim, A. S. N., Kapler, I. V., & Foot, V. (2019). Enhancing the quality of student learning using distributed practice. In J. Dunlosky & K. Rawson (Eds.), *Cambridge handbook of cognition and education* (pp. 550-584). Cambridge, UK and New York, NY: Cambridge University Press.

Wulf, G., & Shea, C. H. (2002). Principles derived from the study of simple skills do not generalize to complex skill learning. *Psychonomic Bulletin and Review*, *9*(2), 185-211. https://doi.org/10.3758/BF03196276

Zarate J. M. (2013). The neural control of singing. *Frontiers in Human Neuroscience*, 7 (237), 1-12. https://doi.org/10.3389/fnhum.2013.00237

Zentner, M., & Strauss, H. (2017). Assessing musical ability quickly and objectively: development and validation of the Short-PROMS and the Mini-PROMS. *Annals of the New York Academy of Sciences*, *1400*(1), 33-45. https://doi.org/10.1111/nyas.13410

Appendix: Sample Script

First session

"Welcome to the song learning study. The first thing we have to do is to go through the consent forms so that you have a clear idea of what the study is about, and can confirm that you are willing to participate. (Distribute consent forms)."

"In this study, you will be presented with a series of recordings which will help you to learn a simple song in two verses. This first session is 35-40 minutes long (for massed condition only: "the second is about 20 minutes after a 10 minute break). Just let me know when you are ready to begin, and we'll get started. If you have any problems, just ask me for help. Do you have any questions?

"Let's start with a warm-up. Have a seat at the monitor. What is your voice type? Sing the phrase after me." Choose appropriate starting pitch and alter if too high or too low. Demonstrate each phrase before they sing it back.

1. Hum up and down an octave. Top note is "original octave note."
2. Sing oo up and down an octave.
3. 1-3-555 on Doo-Doo-Doo etc.
4. Same sequence one half tone higher: humming up and down the octave, then OO, then 1-3-555 on Doo.
5. Continue chromatically four more times with Doo-Doo-Doo until "original octave note" is reached.

Slide 1: {Soprano} Song Study

Slide 2: "This slide gives us a chance to adjust the monitor levels. Please let me know if you would like an adjustment to the volume."

Slide 3: Hi. You are here for a study of song learning. I am going to ask you to practice a short song in two verses, and when you think you know it, to sing it for me from memory. The computer will guide your learning. When you are ready, say "O.K." and I will change the slide.

Melody learning trials:

Slide 4: "You will now hear a piano recording of the song we are asking you to learn. For this presentation, there is no score. Say "OK" when you are ready for the piano recording."

Slide 5: Piano recording of song, two verses:[12]

Slide 6: "Here is the melody, one phrase at a time. Each phrase is repeated. Sing each line after you hear it". **Script: Do you have any questions?**

Slide 7: Phrase one

---

[12] In every case, recordings will follow two metronome clicks in the tempo of the song.

Slide 8 Phrase one again

Slide 9: Phrase two

Slide 10: Phrase two again

Slide 11: Phrase 3

Slide: 12: Phrase 3 again

Slide 13: Phrase 4

Slide 14: Phrase 4 again

Slide 15: This time you will hear the starting note, two clicks, then the piano recording.

Slide 16: Piano recording with score.

Slide 17: "Please sing along with the recording on the syllable "na". There are 3 repetitions in this sequence. Say "OK" when you are ready to continue."

Slides 18-20: Piano score. Piano recording of song with "na", two verses. (Note: if recording does not play, try up arrow and space bar.)

Slide 21: "Try the melody on your own, from the score. You will hear the starting pitch, two clicks, then the first notes of the tune. That is your cue to sing."

Slide 22: Melody cue with score.

If there are no errors, press spacebar to advance.
If there are errors, say: "There are a couple of errors. Let's go back so you can hear the melody again." Press object shape (back to 17) then spacebar. Continue after sequence repetition.

Word study:

Slide 23: Thanks for learning the melody. Now take a few minutes to look at the lyrics.

Slide 24: Song text
It's time for vacation I studied enough
I am sick of my courses the profs are too tough}Courses
I don't have any reason to get out of bed
And the thought of my homework just fills me with dread. }Bed

Slide 25: Song text
My data is done and my phone has no juice
No more internet dating I can't see the use. }Data
All the stress is a pain and I need some release
But my head is exploding so leave me in peace }Stress/head

Song learning trials:

Slide 26: "You will now hear the song with words, all the way through. For this presentation, there is no score."

Slide 27: Voice recording of entire song, both verses.

Slide 28: Here is the song, one line at a time. Sing each line right after you hear it.

Slides 29-44: The song one line at a time repeated (no score).

Slide 45: Great. Now try it two phrases at a time. Please sing after you hear each section.

Slides 46-53: The song two lines at a time repeated (with score).

Slide 54: Listen to the song 3 more times. Sing along with the words. Say "OK" when you are ready to continue.

Slide 55-57: Score and voice recording of entire song, both verses.

Slide 58: "Say "Yes" if you are ready to try the song on your own, from the score. Say "Go back" to hear the song again." **If yes, continue. If no, press object button (return to 54) and spacebar. Allow as many repetitions as requested, but keep a record.**

Slide 59: "I would like you to sing the song from the score. You will hear the opening pitch, then two metronome clicks, then the first notes and words of the song. That is your cue to begin.

Slide 60: Song Cue. Opening pitch, then two metronome clicks, the first three notes and words of the song,

If errors, first time only: "There are a few errors. Let's go back to hear the song again." **then press object button (return to 54).**
If correct: continue. Allow a total of one repetition of the sequence if necessary, then continue even if no criterion reached. Keep a record of any repetition.

Slide 61: "Here are three repetitions of the song with the score. Please sing along with the words. Take as much time as you need with each slide to go over the words, then say "OK" to move on when you are ready."

Slide 62-64: song trials

Slide 65: "When you are ready, please sing the song from memory. You may start again from the beginning once, if you get stuck, but try to keep going, even if you are unsure. Say "Yes" to sing the song from memory or say "Go back" to hear the song again." Object button returns to 61.

Slide 66: Song Cue. Repeat memory test as much as 6 times. Coach as needed. Object button returns to 64. If errors, coach as needed:

Coaching:

The first phrases are about vacation: "It's time for vacation. Why? You need a vacation because have studied enough. Why else? Because you are sick of your courses, and especially because of the tough profs." Now speak the lines: (prompt if necessary: "It's time for vacation…I am sick…")

The next phrases are about bed: You don't want to get out of bed, because your homework is due and you haven't done it. Speak the lines: (prompt if necessary: "I'don't have any… and the thought").

The next part is all about data and dating! Try the lines: (prompt if necessary: "My data is done…no more internet dating)

And the last phrase is about stress that gets to your head: Try the lines: "All the stress is….but my head…)

e.g. "The last section was about stress that got to your head. Listen again, then try again from memory."

Allow 45 minutes or 95% criterion to complete session 1, then say "That was great. Please take a few minutes for a drink of water and to fill out a short questionnaire."

(Concluding remarks depend on experimental condition)

Second session:

At the door:

Thanks for returning for your second session. Just like the first time, I am going to set you up at the computer terminal with headphones and a microphone. We will check sound levels right off the top, to make sure you can hear us and we can hear you. Then I will ask you to sing the song you learned at the first session.

(sound test)

Please let me know when you are ready to start.

Slide 67: Opening screen

Slide 68: I would like you to sing the notes and words of the song you learned at the first session. You will hear the starting pitch, then two clicks, then the first notes of the tune. That is your cue to sing.

Slide 69: You may start again from the beginning once, if you get stuck, but try to keep going, even if you are unsure. Say O.K. when you are ready to continue".

Slide 70: Song cue piano

Slide 71: Sing the song again, this time cued by the notes and the words. You may start again from the beginning once, if you get stuck, but try to keep going, even if you are unsure. Say "OK" when you are ready to continue.

Slide 72: Melody and Word cue

Slide 73: You will now hear the song with words, 3 more times. Please sing along, following the score. When you are ready, say "OK" and the song will play."

74-76: Voice recording of entire song, both verses

Slide 77: I would like you to sing the song from memory, cued by the first notes. You may start again from the beginning once, if you get stuck, but try to keep going, even if you are unsure. Say "OK" when you are ready to continue.

Slide 78: Song Cue piano

Slide 79: I would like you to sing the song from memory, cued by the first notes and words. You may start again from the beginning once, if you get stuck, but try to keep going, even if you are unsure. Say "OK" when you are ready to continue. Thank you. See you in three weeks for your final session.

Slide 80: Song Cue

Final test

At the door

Thanks for returning for your final session. Just like the other times, I am going to set you up at the computer terminal with headphones and a microphone. We will check sound levels right off the top, to make sure you can hear us and we can hear you. Then I will ask you to sing the song you learned at the first session.

Slide 81: Preset on the monitor:

Text display: Opening:

Slide 82: I would like you to sing the song you learned at the previous sessions cued by the first notes. You may start again from the beginning once, if you get stuck, but try to keep going, even if you are unsure. Say "OK" when you are ready to continue.

Slide 83: Song Cue. Pitch, two metronome clicks, then the first two notes of the song.

Slide 84: Sing the song again, this time cued by the notes and the words. You may start again from the beginning once, if you get stuck, but try to keep going, even if you are unsure. Say "OK" when you are ready to continue.

Slide 85: Song Cue. Pitch, two metronome clicks, then the first three notes and words of the song.

Slide 86: Listen to the song again. Sing along with the words.

Slide 87: Song with score.

Slide 88: Please try the song from memory.

Slide 89: Song cue.

Return to 86 until criterion is reached.