Manuscript accepted for publication in Memory

The impact of music on learning and consolidation of novel words

Jakke Tamminen[a], Kathleen Rastle[a], Jess Darby[a], Rebecca Lucas[a], and Victoria J. Williamson[b]

[a]Department of Psychology, Royal Holloway, University of London, Egham, TW20 0EX, United Kingdom

[b]Department of Music, University of Sheffield, Sheffield, S3 7RD, United Kingdom

Author note

Jakke Tamminen, jakke.tamminen@gmail.com; Kathleen Rastle, kathy.rastle@rhul.ac.uk; Jess Darby, Jess.Darby.2009@live.rhul.ac.uk; Rebecca Lucas, Rebecca.Lucas@roehampton.ac.uk; Victoria Williamson, v.williamson@sheffield.ac.uk.

Address correspondence to: Jakke Tamminen, Department of Psychology, Royal Holloway, University of London, Egham, TW20 0EX, United Kingdom. E-mail: jakke.tamminen@gmail.com. Tel: +44(0)1784 414635.

Abstract

Music can be a powerful mnemonic device, as shown by a body of literature demonstrating that listening to text sung to a familiar melody results in better memory for the words compared to conditions where they are spoken. Furthermore, patients with a range of memory impairments appear to be able to form new declarative memories when they are encoded in the form of lyrics in a song, while unable to remember similar materials after hearing them in the spoken modality. Whether music facilitates the acquisition of completely new information, such as new vocabulary, remains unknown. Here we report three experiments in which adult participants learned novel words in the spoken or sung modality. While we found no benefit of musical presentation on free recall or recognition memory of novel words, novel words learned in the sung modality were more strongly integrated in the mental lexicon compared to words learned in the spoken modality. This advantage for the sung words was only present when the training melody was familiar. The impact of musical presentation on learning therefore appears to extend beyond episodic memory and can be reflected in the emergence and properties of new lexical representations.

Keywords: Word learning, memory, music, memory consolidation, lexicalisation

The idea that music is a useful tool for committing new information to memory is widespread and popular. Songs are used to help children and adults learn a multitude of ideas from the letters and sounds of the alphabet to the basic principles of physics (e.g., Dickson & Grant, 2003). There is substantial empirical evidence that music can be an effective mnemonic aid in memorising lyrics or word lists (e.g., Wallace, 1994). However, most existing work has been dedicated to examining the impact of music in tasks that rely largely (but not always exclusively) on episodic memory. In the present work we examine the impact of music on the acquisition of completely novel stimuli. In addition, this allows us for the first time to investigate the impact of music on key memory processes that are known to be integral to word learning (see e.g., Davis & Gaskell, 2009). Specifically, we employ a word learning paradigm that allows us to determine whether new words learned through listening to singing become integrated in the existing mental lexicon in the same way as words learned in the spoken modality. We also seek to establish when this integration takes place; previous research has shown that the integration of spoken novel words requires a period of memory consolidation (ideally a night of sleep; Dumay & Gaskell, 2007). We ask whether music can speed up or strengthen this integration process. We start with a brief overview of existing studies examining the role of music in declarative memory before turning to the question of music and word learning in greater detail.

Impact of music on declarative memory

Most existing research compares explicit memory for verbal material presented in a musical vs. non-musical manner using mainly episodic memory tasks such as free recall or recognition memory. For example, Wallace (1994) asked adult participants to memorise the lyrics to a ballad with the words presented either in the spoken or sung modality. Participants were asked to recall the words verbatim both during the training, which consisted of repeated presentations of the ballad, and in a delayed test 20 minutes later. Recall accuracy was higher

in the sung condition during training and continued to be higher in the delayed test. Similar findings on the superiority of learning lyrics in the sung modality as measured by verbatim recall were reported by Calvert and Tart (1993), Kilgour, Jakobson, and Cuddy (2000), and McElhinney and Annett (1996). It is worth highlighting that all of these authors, like Wallace (1994), found that the benefit of the sung modality increased as familiarity with the melody increased. In one case the sung benefit was wholly restricted to conditions in which the song was heard multiple times as opposed to just once (Calvert & Tart, 1993).

While a number of studies have shown that music may benefit verbal memory, the evidence is not entirely unequivocal. For example, Racette and Peretz (2007) manipulated both the modality of the stimulus presentation, and the modality in which participants were required to recall the text of newly learned songs but failed to find any benefit of sung presentation over spoken presentation at encoding. Although this finding implies that more research is needed to elucidate the circumstances under which musical presentation has a benefit on verbal memory, it is worth noting that the spoken condition of the Racette and Peretz study, presented a background melody while participants heard the spoken text. Therefore this condition was not purely non-musical, and the background music may have increased recall rates (see Kang & Williamson, 2014, for evidence that background music may have a beneficial effect on memory), thus possibly obscuring any difference between spoken and sung conditions.

Impact of music on semantic memory in patients with memory impairments

Recent work on patients suffering from memory impairments provides insight into the potential neural basis for the benefit of music in memory. A number of studies have shown that patients with Alzheimer's disease (AD) recognise lyrics that they heard sung more reliably than lyrics heard in the spoken modality (Simmons-Stern et al., 2010). They are also

able to retain more of the semantic content of the lyrics learned in the sung modality (Simmons-Stern et al., 2012). Simmons-Stern and colleagues (2010, 2012) have suggested that the benefit of music on memory is due to the more diversified neural encoding of musical stimuli compared to non-musical stimuli. AD patients typically show cortical and medial temporal lobe (MTL, including the hippocampus) atrophy, a pattern associated with impaired episodic learning. Music processing has been shown to engage a broad and complex neural network encompassing cortical and subcortical areas outside of the MTL (e.g., Koelsch, 2011; Mueller et al., 2015; Peretz & Zatorre, 2005) and it may be that the engagement of such broad networks allows robust encoding of musical sounds in memory (including sung words) even for patients suffering from MTL-related atrophy. The robustness of such musical memories is supported by work by Moussard and colleagues (2012, 2014) and Palisson et al. (2015) who in AD patients and matched controls have shown a benefit of the sung modality even several weeks after learning. Similar results have also been reported in patients with multiple sclerosis (Thaut, Peterson, McIntosh & Hoemberg, 2014).

Further evidence for the notion that musical presentation can compensate for deficits in the function of the MTL and the hippocampus come from studies on amnesic patients with damage to these areas. Baur et al. (2000) reported a case study of an amnesic patient with severe impairment of declarative memory in general, but who was nonetheless able to learn the titles of songs she was learning to play. Haslam and Cook (2002) reported two amnesic patients who were asked to discriminate between lyrics they had been trained on before the test and lyrics that remained untrained. The patients were successful, but only if the lyrics were heard in the sung modality during training rather than in the spoken modality. In addition, these patients had more accurate memory for the semantic content of the lyrics if they had been trained in the sung modality.

Word learning as a window into memory processes

The literature reviewed above suggests that declarative memory benefits from musical presentation of verbal materials in both normal adults and in patients suffering from memory impairments. The patient literature also suggests that this benefit may be due to music recruiting a broader network of brain areas during encoding than non-musical presentation. In the current series of experiments we seek to address two as yet unanswered questions. Firstly, while music may assist memory for already familiar verbal stimuli such as lyrics or word lists, it is completely unknown what role music plays in learning *new* information and its integration in semantic memory. To address this issue we teach adult participants novel words in sung or spoken modalities, and test for the first time the impact of varying musical presentation on both explicit and implicit memory of the newly learned stimuli. While we teach participants new words in their own language, there is encouraging recent experimental evidence from second language learning suggesting that music may help memory for foreign words. When Spanish-speaking children were required to memorise a passage of text in English (an unfamiliar language to them), significantly better performance was seen in verbatim recall, pronunciation, and translation when the text was learned as a song compared to speech (Good, Russo & Sullivan, 2015, see also Medina, 1993).

Secondly, we employ a behavioural technique to evaluate the hypothesis that singing may allow the encoding of information in a broad cortical network with little involvement of the MTL and the hippocampus. Specifically, we employ a word-learning paradigm that offers a behavioural diagnostic of MTL-dependent early lexical representations and MTL-independent consolidated lexical representations, as put forward by Davis and Gaskell (2009).

Davis and Gaskell (2009) presented a theory of the cognitive and neural processes involved in learning new spoken words, based on the general principles of the Complementary Learning Systems (CLS; McClelland et al., 1995) account of memory. This theory argues that newly-learned spoken words are initially encoded by the hippocampus and related MTL structures. After a period of offline memory consolidation (that seems to involve neural processes specific to sleep; Tamminen et al., 2010), these novel words gradually become represented in neocortical areas. The theory also proposes that these two learning systems rely on different architectures. The fast learning hippocampus codes information in non-overlapping, distinct representations, while the slow learning neocortex represents information in overlapping representations, hence allowing the full integration of new memories with existing knowledge as well as the discovery of shared patterns of information across a large number of memories (Tamminen et al., 2015).

Davis and Gaskell (2009) argued that the large body of evidence that exists in the domain of spoken word learning fits in elegantly with predictions made by the CLS account. For example, Gaskell and Dumay (2003) trained participants on novel spoken words (e.g., *dolphik*) that share a large part of their phonological onset with existing words (e.g., *dolphin*). Because the recognition time of spoken words depends largely on the number of phonologically overlapping competitors (Marslen-Wilson, 1987), Gaskell and Dumay (2003) hypothesised that learning a new competitor such as *dolphik* should result in delayed recognition of the base word *dolphin*. Sure enough they reported slower recognition times to base words for which new competitors had been taught, but critically, this was found only if participants were tested at least 24 hours after training, and not immediately after training. The authors argued that this delayed lexical competition effect was due to a need for memory consolidation to operate before the new lexical representations could be integrated in the mental lexicon. A study by Davis et al. (2009) using fMRI confirmed that the slow

emergence of the lexical competition effect was associated with a shift from early hippocampal involvement to post-consolidation neocortical representation.

Another notable feature of the Gaskell and Dumay (2003) experiments and the many studies that later replicated the findings (e.g., Dumay & Gaskell, 2007; Tamminen & Gaskell, 2008; Davis et al., 2009) was that explicit measures of memory for the novel words, recognition memory and free recall, were all very high immediately after training. This led Davis and Gaskell (2009) to suggest that newly learned spoken words form highly accurate representations mediated by the hippocampus immediately after training. However, because the hippocampus employs distinct, non-overlapping representations, the new lexical representations are not integrated with the existing mental lexicon. It is only over the course of memory consolidation that neocortical, overlapping representations are formed, allowing lexical competition effects to emerge.

Given the relatively well understood processes involved in spoken word learning, we suggest that the above paradigm can be used to gain a better understanding of how variables such as music impact on learning and memory. If musical presentation leads to better learning and stronger memories than non-musical presentation, we would expect novel words learned in the sung modality to have an advantage over words learned in the spoken modality in a test of explicit memory. Such a prediction is supported by the literature on music and declarative memory reviewed earlier (Calvert & Tart, 1993; Kilgour, Jakobson & Cuddy, 2000; Wallace, 1994). In addition, we can use the emergence of the lexical competition effect as a marker of the point in time where the newly learned words become represented in broad neocortical networks. Previous research in the spoken modality shows that this requires at least one night of sleep after training (e.g., Dumay & Gaskell, 2007). However, if it is the case that in normal adult learners, as in patients with memory impairments, learning in the sung modality engages broader extra-MTL areas of the brain, we might observe lexical

competition effects earlier when words are learned in the sung modality compared to the spoken modality, and the effects might be larger in magnitude, indicating stronger integration in the mental lexicon. Here we report three experiments in which participants learned novel spoken words (e.g., *dolphik*) either in the spoken or sung modality. We tested explicit memory of the words as well as the lexical competition effects immediately after training, one day after training, and one week after training.

Experiment 1: Learning new spoken words

In Experiment 1 we trained participants on 32 novel words in the *spoken* modality using a phoneme monitoring task. In addition to providing information about recall levels and the magnitude and time course of the lexical competition effect in the spoken training modality, this experiment was also an important test to establish that we could obtain robust learning and lexical competition effects using a training task modelled after the typical phoneme monitoring task used in previous word learning studies but which had been modified to accommodate musical presentation. In the typical version of this task (e.g., Gaskell & Dumay, 2003; Dumay & Gaskell, 2007; Tamminen & Gaskell, 2008) participants listen to novel words through the headphones and after presentation of each word are asked to decide whether a pre-determined target phoneme was present or absent in the word. In the present series of experiments we needed to present the novel words in the format of a song (in Experiments 2 and 3), therefore words could not be presented individually one-by-one but rather had to be heard in a continuous manner. Therefore we created a go/no-go version of the task where the words were presented in lists of continuous strings with only a short gap of 500ms separating each word. The task is described in the Methods section.

To measure explicit recall we adopted the two tasks used by Tamminen et al. (2010): free recall and old-new categorisation. Free recall rates in word learning studies tend to be low and may underestimate participants' knowledge of the words. Therefore this task is often complemented by recognition memory tasks such as the old-new categorisation task. This task measures participants' knowledge of the phonological configuration of the newly learned words, and in addition provides a reaction time measure of access to newly created lexical representations, thus providing a method of comparing explicit memory in the spoken and sung conditions.

To measure lexical competition, we followed Gaskell and Dumay (2003) and others (e.g., Dumay & Gaskell, 2012) and chose the pause detection task. In this task participants are asked to listen to spoken words and to monitor for words that have a short pause embedded in the speech. Mattys and Clark (2002) have shown that the time it takes to make a pause decision reflects the amount of lexical activity at the time. Therefore words with many competitors are associated with slower pause detection times than words with fewer competitors. In our test sessions we asked participants to listen to base words that had a trained new competitor (e.g., *dolphin*) and base words that had no newly learned competitors (e.g., *falcon*). We predicted that a difference in pause detection times to the base words should emerge once the novel word (e.g., *dolphik*) had been integrated in the mental lexicon, with slower pause detection times observed to base words with a new competitor.

Method

Participants. Thirty-nine native English-speaking participants completed the study (25 female, 5 left-handed, mean age = 21). None reported suffering from language or hearing disorders. All were students or staff at Royal Holloway, University of London, and were paid for their participation. All participants were screened prior to taking part to ensure they were

native speakers of British English and non-musicians, defined as someone who has not undertaken any musical training outside of their school curriculum and is not currently training on a musical instrument or voice.

Materials. 64 familiar monomorphemic base words (e.g., *dolphin*) and novel word pairs derived from each base word (e.g., *dolphik* and *dolphis*) were selected from the pool of stimuli used by Gagnepain et al. (2012). One of the two novel words in each pair was used for training, and the other one was used as a foil in the old-new categorisation task. All base words were bisyllabic and 4-8 phonemes long (M=5.78). CELEX frequencies (Baayen et al., 1993) of the base words ranged from 1 to 76 occurrences per million (M=10.62). 59 of the 64 base words had an early uniqueness point (before the final vowel). Novel words were derived from base words by changing one or two final phonemes. Five of the base words had a later uniqueness point: novel words were derived from these words by adding a phoneme to the end (e.g., *widow – widowl*). These 64 stimulus triplets were divided into two lists, one to be used in the trained condition and the other to remain untrained. The two lists were matched in frequency, number of phonemes, and uniqueness point, and were counterbalanced across participants so that both lists were used in the trained and untrained conditions.

All spoken stimuli were recorded in a soundproof booth by one of the authors (VJW) who is a native speaker of British English. Two tokens of each novel word was recorded; one to be used in the training task and another one to be used in the old-new categorisation task, in order to prevent participants making the old-new categorisation response purely based on familiarity with the acoustic form of the stimulus.

Procedure. Participants first completed the training session where they were familiarised with the novel words in a phoneme monitoring task. This was followed by the first test session conducted immediately following training, where participants carried out a

pause detection task, a free recall task, and an old-new categorisation task. The test session was repeated one day after training, and once more one week after training. All tasks (except free recall) were carried out on computers running DMDX (Forster & Forster, 2003), with standard keyboards used for response collection in the training phase, and button boxes in the test phase.

Training session. The training session consisted of a phoneme monitoring task, a modified version of that used in previous spoken word learning studies (e.g., Tamminen & Gaskell, 2008). In the current version participants listened to an uninterrupted list of 32 novel spoken words, with each word separated by a gap of 500ms of silence, and were asked to press a response button every time they heard a word that contained a pre-determined target phoneme. The list was presented 36 times, thus giving 36 exposures to each novel word. The order of the words within the list was randomised, but the same order was used in each repetition of the list. To avoid using only one order throughout the experiment, 12 different random orders were created (for each of the two sets of 32 novel words) and each order was used roughly an equal number of times across all participants. Before each presentation of the list, participants were informed of the target phoneme they were to monitor during that particular presentation. The target phoneme remained on the screen for the duration of the presentation of the list. The six target phonemes included /p/, /d/, /m/, /t/, /n/, and /s/. The training session lasted about 45 minutes.

Test session. The three test tasks were carried out in fixed order. The test session started with the pause detection task. Participants heard a spoken word through the headphones, and had to decide as quickly as possible whether it contained a 200ms pause by pressing a "Yes" or a "No" button on the button box. In this task participants heard all 32 base words (e.g., *dolphin*) for which a new competitor (e.g., *dolphik*) had been trained, 32 control base words for which no new competitor had been trained (e.g., *falcon*), and 128 filler

words. The filler words were monomorphemic, bisyllabic words, ranging in CELEX frequency from 1 to 77 (M=10.80), and ranging in length from 4 to 8 phonemes (M=5.84). Thus the fillers were closely matched to the base words in these key properties. Half of the base words had a pause inserted, while half did not. The assignment of base words into the pause-present and pause-absent conditions was counterbalanced across participants. Following Gaskell and Dumay (2003), in the pause-present base words the pause was always inserted before the final vowel (e.g., *dolph_ik*). Half of the filler items also contained a pause but here it could occur in any position of the word. Order of presentation of the stimuli was newly randomised for each participant in each session. Reaction times were measured from the onset of the spoken word (although at the analysis stage these were adjusted to measure RTs from the onset of the pause in pause-present trials, and in pause-absent trials from the point where the pause would have been inserted, again following Gaskell and Dumay, 2003), and the response deadline was set at 3000ms from the onset of the word.

The free recall task followed the pause detection task. Participants were given three minutes to recall as many of the novel words as possible in any order. They were asked to say the words aloud as they recalled them, and responses were recorded for later scoring.

Finally, in the old-new categorisation task participants were presented auditorily with novel words and their foils. The task was to indicate with a key press on the button box whether the word was a trained novel word or a similar-sounding foil. In the first test session, only half of the novel words (and their foils) were presented. The second session included all 32 novel words (and foils). This allowed us to restrict analysis in the second session to only those items that had not been experienced in the first session, to avoid repetition effects. The assignment of items in the first session to the presented and withheld conditions was counterbalanced across participants. In all test sessions the presentation order of the stimuli was pseudo-randomised with the constraints that at least four trials had to intervene the

presentation of a novel word and its foil, and that half of the novel words preceded its foil and half followed it. Four unique orders were created for each session and used an equal number of times across participants. RTs were measured from the onset of the word, with a response deadline set at 4000ms.

Results

Reaction time data were analysed using mixed-effects modelling (Baayen, Davidson, & Bates, 2008) in R using the *lme4* package. This decision allowed us to include participants and items simultaneously in the same model. Random effects structure was always determined by comparing a series of models with gradually simplifying structure, thus preserving those factors that contributed significantly to the model fit. Likelihood ratio tests were carried out to evaluate the significance of each fixed effect by comparing a model that includes the effect to an identical model that does not include the effect (Barr, Levy, Scheepers, & Tily, 2013). Accuracy data were analysed according to the same strategy using logistic mixed-effects models. Here, the p-values are reported based on the Wald Z statistic for each fixed effect (Jaeger, 2008).

Training. To ensure that participants were attending to the training task and that the novel words were intelligible, we examined accuracy data in the phoneme monitoring task. Recall that the task consisted of a continuous presentation of a list of novel words (with the list repeated a total of 36 times) rather than presentation of distinct trials, with a response required only when the participant detected a word where the target was present. We therefore counted the number of target-present responses made each time the list was listened to and compared it to the number of novel words in the list where the given target was present (i.e. the correct number of target-present responses). We then calculated the proportion by which the number of target-present words was misestimated (e.g., if the number of words

with a target present was 8 and the participant made 6 target-present responses, they misestimated by 25%). Calculated across the 36 presentations of the list and across all participants, participants made on average a misestimate of 23%. This value consists of both misses (underestimation) and false alarms (overestimation).

Pause detection. One participant's data were lost in the pause detection task due to experimenter error. Following Dumay and Gaskell (2012), pause detection data were collapsed over pause-present and pause-absent trials. Erroneous responses were removed, as were extremely long or short RTs (above 2000 ms or below 150 ms; 0.3% of the data). The data were then log-transformed to better meet the assumption of normality and to reduce the effect of remaining outliers. Data in all tables and figures are retransformed. Training (trained competitor vs. no trained competitor) and test session (first vs. second vs. third) were included as fixed factors. By-subjects random slopes for test session were included, as they significantly improved the model fit. The factor of training contributed significantly to the model, $\chi^2(1)=12.68$, $p<.001$, but test session did not. Importantly, the interaction between the two factors was significant, $\chi^2(2)=8.18$, $p=.02$. This interaction reflected the fact that while there was no training effect observed in the first test session, $\chi^2(1)=0.07$, $p=.79$, there was a significant training effect in the second session, $\chi^2(1)=9.33$, $p=.002$, and in the third session, $\chi^2(1)=10.78$, $p=.001$. The pause detection data are summarised in Table 1, and the magnitude of the lexical competition effect at each test session in Figure 1.

-- Insert Table 1 about here –

Accuracy rates in the pause detection task are presented in Table 1. Training (trained competitor vs. no trained competitor) and test session (first vs. second vs. third) were included as fixed factors. No random slopes were included as they did not significantly

improve the model fit. No significant effects of training, test session, or an interaction between the two were found.

-- Insert Figure 1 about here –

Free recall. Free recall data (Figure 2) were analysed using a logistic mixed-effects model with test session as a fixed factor. No random slopes were included. The analysis revealed a significant main effect of test session, $\chi^2(2)=24.04$, $p<.001$. A comparison of free recall rates across the three days showed that while there was no significant difference in recall rates between sessions 1 and 2, recall in session 3 was significantly higher than recall in session 1, $z=4.62, p<.001$, or session 2, $z=3.76, p<.001$.

-- Insert Figure 2 about here –

Old-new categorisation. Following Tamminen et al. (2010), in the RT analysis erroneous responses and extremely long or short RTs (above 3000 ms or below 500 ms; 0.2% of the data) were removed. The data are summarised in Figure 3. Test session (first vs. second vs. third) was included as a fixed factor. By-subjects random slopes for the effect of test session were retained. The main effect of session was significant, $\chi^2(2)=8.89$, $p=.01$. Pairwise comparisons of the three sessions showed a significant difference between sessions 2 and 3, $\chi^2(1)=8.59$, $p=.003$, but no other contrasts were significant.

Accuracy in the old-new categorisation task was analysed by calculating signal detection measures (d') in order to take into account response bias. Memory of novel words was evaluated by calculating the difference between z-transformed proportion of accurate "yes" responses to trained novel words (hits) and incorrect "yes" responses to foils (false alarms). These data are presented in Figure 3. Since item-level data are not available when

analysing d' values, we used analyses of variance (ANOVAs). An ANOVA with test session as a within-participants factor showed no significant main effect of test session ($p=.21$).

-- Insert Figure 3 about here –

Discussion

The results of Experiment 1 replicated the typical pattern of slowly emerging lexical competition effects following repeated presentation of novel words. There was no competition effect immediately after training, but a reliable effect was seen the following day and a week later. This suggests that when trained in the spoken modality, novel words became integrated with the existing lexicon only after a 24-hour consolidation opportunity. This replication was consistent with the literature in spite of the modifications made to the typical phoneme monitoring training task, namely the adoption of word list stimuli and a go/no go training format.

Free recall rates were also comparable to previous studies and increased over time (e.g., Tamminen et al., 2010). The increase observed here was likely due to practice with the task, and extra exposures to trained novel words gained over the course of testing with the old-new categorisation task. Results of the old-new categorisation task were also consistent with data reported in previous studies: RTs in this task got faster over time as a function of practice, while accuracy remained relatively stable over time.

In sum, Experiment 1 was successful in establishing that the typical pattern of lexical competition and memory effects can be obtained with our new go/no-go training task. In the next two experiments we repeated these tasks but presented our novel words in the sung modality, and compare these data to the present spoken modality baseline.

Experiment 2: Learning sung words with unfamiliar melody

In Experiment 2 participants learned the same novel words as in Experiment 1, and were tested on the same tasks and using the same stimuli as in Experiment 1. The major difference was that in this experiment the novel words in the training sessions were presented in the sung modality. The training task and the number of exposures was the same as in Experiment 1, therefore the only difference in training was in the modality.

Method

Participants. Thirty-nine non-musician native English-speaking participants completed the study (25 female, 5 left-handed, 1 ambidextrous, mean age = 21). None reported suffering from disorders affecting language or hearing. All were students or staff at Royal Holloway, University of London, were paid for their participation, and none had taken part in Experiment 1.

Materials. The same familiar and novel word stimuli were used as in Experiment 1. The melodies that formed the basis of the sung stimuli were selected from a hymn database assembled from a Church of England traditional hymnal (Nicholson, Knight, Dykes, & Bower, 1950). This hymn database was developed for use alongside a computational model of melodic expectation, based on information theory and statistical learning principles (Pearce, 2005; Pearce & Wiggins, 2006; Pearce et al., 2010). The hymn melodies have previously been used to examine musical understanding in a wide range of populations including individuals with specific music processing difficulties (congenital amusia; Omigie, Pearce, Williamson & Stewart, 2013) so are deemed to be suitable for the present non-musician participants.

Melodies for the present study were selected from the hymnal database according to their length, so that each note could be paired with a novel word from the Experiment 1 lists. We selected six melodies from the database that each comprised 32 isochronous notes. The melodies were transposed in manuscript form from their original database keys to four tonalities that were within the range of the singer (three melodies in C Major, one in G major, one in F major and one in D flat major).

The sung stimuli were recorded in a soundproof booth by the same speaker who recorded the spoken stimuli (VJW). The singer recorded each 32 novel word list using each of the six melodies. An example of one of the stimuli, melody and word list combined, can be seen in Figure 4. Like in Experiment 1, the recordings were later edited so that there was a 500ms gap of silence between the offset and onset of each word.

-- Insert Figure 4 about here –

Procedure. Both the training and test sessions and the tasks carried out in these sessions were identical to Experiment 1 except that the training involved sung rather than spoken stimuli, as described above.

Results

Training. Accuracy in the phoneme monitoring training task was calculated in the same way as in Experiment 1. Training data from one participant was lost due to equipment failure. Participants misestimated the number of target-present words on average by 27%.

Pause detection. The pause detection data are presented in Table 1 and Figure 1[1]. As before, erroneous responses and extremely long or short RTs (above 2000 ms or below

[1] One base word (*guitar*) was removed from the pause detection analysis in this experiment and in Experiment 3 because the corresponding novel word (*guitas*) was stressed incorrectly in the sung recordings.

150 ms; 0.2% of the data) were removed and the RTs log-transformed. Training (trained competitor vs. no trained competitor) and test session (first vs. second vs. third) were included as fixed factors. By-subjects random slopes for test session were included, as they significantly improved the model fit. The factor of training contributed significantly to the model, $\chi^2(1)=19.92$, $p<.001$, as did test session, $\chi^2(2)=13.82$, $p=.001$. The interaction between training and test session too was significant, $\chi^2(2)=7.79$, $p=.02$. This interaction reflected the fact that while there was no training effect observed in the first test session, $\chi^2(1)=1.26$, $p=.26$, there was a significant training effect in the second session, $\chi^2(1)=4.00$, $p=.046$, and in the third session, $\chi^2(1)=19.21$, $p<.001$.

Accuracy rates in the pause detection task are presented in Table 1 and were analysed as before. Training (trained competitor vs. no trained competitor) and test session (first vs. second vs. third) were included as fixed factors. No random slopes were included as they did not significantly improve the model fit. No significant main effects of training were found but a significant effect of test session did emerge, $\chi^2(2)=6.28$, $p=.04$. This reflected a significant difference in accuracy between sessions 1 and 3, $z=2.35$,$p=.02$, no other contrasts were significant.

Free recall. No random slopes were included in the model for free recall data. The analysis revealed a significant main effect of test session, $\chi^2(2)=35.23$, $p<.001$. A comparison of free recall rates across the three days found no significant difference in recall rates between sessions 1 and 2, while recall in session 3 was significantly higher than recall in session 1, $z=5.23$,$p<.001$, or session 2, $z=4.97$,$p<.001$ (Figure 2).

Old-new categorisation. Erroneous responses and extremely long or short RTs (above 3000 ms or below 500 ms; 0.3% of the data) were removed. Test session (first vs. second vs. third) was included as a fixed factor. By-subjects random slopes for the effect of

session were retained. The main effect of session was significant, $\chi^2(2)=8.71$, p=.01. Pairwise comparisons of the three sessions showed a significant difference between sessions 2 and 3, $\chi^2(1)=6.94$, p=.008 and sessions 1 and 3, $\chi^2(1)=5.43$, p=.02, but no other contrasts were significant (Figure 3).

Accuracy data are presented in Figure 3. An ANOVA with test session as a within-participants factor showed no significant main effect of test session (p=.60).

Comparison across Experiments 1 and 2. To examine differences between the lexical competition effect observed in Experiment 1 and Experiment 2, we combined pause detection data from the two experiments. Data were trimmed in the same manner as before. We entered training (trained competitor vs. no trained competitor), test session (first vs. second vs. third), and experiment (spoken vs. sung) as fixed factors. By-subjects random slopes for training were also included. No three-way interaction was found, suggesting that both experiments showed a similar pattern of data regarding the emergence over time of lexical competition effects. We then tested the interaction between training and experiment, so see if the lexical competition effect was larger in Experiment 2. This interaction was not significant. To confirm this and the lack of differences between the experiments in the time course of the effect of training, we analysed each test session separately. There was no interaction between the effect of training and experiment in any of the three sessions (all ps>.05). As this interaction measures the difference in the lexical competition effect across the two experiments, this result indicates that the lexical competition effect was indeed statistically identical in both experiments in all three test sessions. No significant interactions with experiment were observed in the analysis of the pause detection accuracy data either.

Free recall data were analysed in a similar manner, with test session and experiment entered as fixed factors. Items-specific slopes were entered for the effect of experiment. We

observed no interaction between experiment and test session, no main effect of experiment, and, consistent with the individual analyses of Experiments 1 and 2, a significant main effect of test session, $\chi^2(2)=58.28$, $p<.001$.

Old-new categorisation RTs were trimmed in the same way as in the main analysis. We entered experiment and test session as fixed factors. Subject and item-specific slopes for the effect of test session were retained. We found no significant interaction between the two fixed factors, no main effect of experiment, and, consistent with the main analyses, a significant main effect of test session, $\chi^2(2)=16.19$, $p<.001$. Accuracy data were analysed with an ANOVA with test session as a within-participants factor, and experiment as a between-participants factor. We observed no interaction between the two factors, no main effect of test session, and a significant main effect of experiment, $F(1,76)=5.12$, $p=.03$.

To compare training performance across the two experiment, we calculated independent-samples t-test. We found no significant difference between the two experiments ($p=.18$), suggesting that participants in both experiments attended to the training equally well, and that the novel words in both experiments were equally intelligible.

Discussion

Experiment 2 showed that novel words learned in the sung modality do become integrated in the mental lexicon, and that this integration occurs over the same time course as words learned in the spoken modality: that is, we found no evidence for lexical competition immediately after training, but a robust effect emerged one day after training and remained significant one week later. This finding fails to support the hypothesis that music might accelerate lexical integration or result in stronger integration effects.

In the free recall data, Experiments 1 and 2 were statistically indistinguishable, suggesting that in the domain of explicit memory we see no benefit for the musical presentation of novel words. In the old-new categorisation task participants were significantly more accurate in Experiment 1 in which the training was in the spoken modality. This accuracy difference is likely to reflect the fact that in Experiment 2, unlike in Experiment 1, there was a sung vs. spoken modality mismatch across the training and testing. Training in Experiment 2 was in the sung modality, but in the test session the novel words were encountered in the spoken modality. Earlier research has established that such training-test modality mismatch typically results in impaired recognition memory accuracy (see Brown & Gaskell, 2014, for a similar mismatch effect when the identity of the speaker is varied between training and test). We can therefore conclude that, consistent with the free recall data, musical presentation failed to benefit recognition memory by neither attenuating nor abolishing the standard mismatch effect. It would be informative to examine the impact of music in the absence of a modality mismatch effect, however the current experiments were designed to investigate effects of musical training on typical spoken word recognition processes and a systematic manipulation of testing modality is therefore beyond the scope of the current set of experiments.

While Experiment 2 did not show evidence for a musical benefit to learning or memory, it would be premature to reject our hypotheses based on these data. As we outlined in the Introduction, many have argued that musical presentation may only provide an advantage if the melody used in the training phase is familiar to the participant (i.e. is repeated; Calvert & Tart, 1993). In Experiment 2 we used church hymns that, while generally familiar in terms of their use of typical Western tonal structures, were unlikely to be individually familiar to the majority of people (especially without their lyrical content). In

Experiment 3 we used the same melodies, but made them familiar to each participant prior to the learning session.

Experiment 3: Learning sung words with familiar melody

Experiment 3 was identical to Experiment 2 in all respects except that participants were familiarised with the melodies used in the novel word training session. The familiarisation occurred over a period of one week preceding the novel word training sessions, and required participants to listen to an instrumental version of one of the hymn melodies several times a day. The effectiveness of this familiarisation was ensured by testing participants' memory of the melody before moving on to the word-learning phase of the experiment. Because the melodies were initially novel and participants only became accustomed to their assigned melody over the course of the familiarisation phase, we were able to ensure relatively equal levels of familiarity. We hypothesised that if the benefit of musical presentation on memory relies on or is significantly enhanced by the familiarity of a melody, we should now observe a difference between the results obtained in Experiment 1 and the current experiment.

Method

Participants. Thirty-nine non-musician native English-speaking participants completed the study (26 female, 5 left-handed, mean age = 20). None reported suffering from language or hearing disorders. All were students or staff at Royal Holloway, University of

London, were paid for their participation, and none had taken part in either Experiment 1 or 2[2].

Materials. The same familiar and novel word stimuli were used as in Experiments 1 and 2, and the same isochronous melodies as in Experiment 2. We created instrumental versions of the melodies to be used in the melody familiarisation phase. The instrumental form of the melodies took the same form as in Omigie et al. (2013): individual notes were created using an electronic piano sound from a MIDI synthesizer before being converted to wav files. The melodies were recorded in an isochronous manner in line with the original hymnal database stimuli, so that each note had the same duration of 1000ms and constant amplitude with no gap of silence between the notes.

Procedure. A week before attending the training session each participant was given the instrumental version of the melody assigned to him or her in an mp3 file format. Participants were asked to listen to their assigned melody twice a minimum of three times every day before the training session, thus resulting in at least 42 exposures over seven days. They were also asked to keep a log of each time they listened to the melody, and to show this log to the experimenter at the beginning of the training session.

To ensure that all participants were familiar with the melody, they were asked to complete a melody memory test upon arrival in the lab for the word-learning phase. The test consisted of eight trials in which an extract of the melody, from the beginning of the melody to a probe point, was played through headphones. The task was to indicate with an untimed key press whether the last note (i.e. the probe) of the extract was correct or incorrect. The

[2] A reviewer queried our choice of a between-participants approach. While the benefits of using the same participants in each condition are significant, we chose to use different participants in the three experiments to avoid the possibility of participants changing their learning strategy over the course of the experiment based on their experience with the different training conditions. This design also solved the problem of having to find a very large stimulus set of base words while controlling their linguistic properties.

eight probe points in the melody were distributed identically in all melodies (at notes 6, 10, 13, 18, 22, 24, 27, and 30).

Half the above points (notes 6, 18, 22, and 27) were altered to a false probe meaning participants heard a note not previously played in that position during the familiarisation phase. False probe tones were selected to have an equivalent information content (IC) level, calculated using predictions from melodic expectation modelling (Pearce & Wiggins, 2006). As such, false probes could be said to be 'as expected' as the original tones in terms of the typical progression of a melody from the hymnal database. In selecting the false probes we chose a note as close as possible to the original, with a minimum of two semitones distance and an equivalent IC level. Melodies with the new false probe tones were recorded using the same protocol as the original instrumental melodies. This procedure reduced the chance that false probes could be identified on the basis that they did not 'fit' as well with the melody as the note heard during the familiarisation phase.

Trials were presented in a fixed order, ascending from the earliest to the latest probe point. Participants were required to achieve an overall success rate of at least 75% in this test to proceed with the experiment. Three participants who failed to reach this criterion were thanked for their time and dismissed.

The training and test sessions and the remaining tasks carried out in these sessions were identical to Experiments 1 and 2.

Results

Training. Accuracy in the phoneme monitoring training task was calculated in the same way as in Experiments 1 and 2. Training data from two participants were lost due to

equipment failure. Participants misestimated the number of target-present words on average by 25%.

Pause detection. The pause detection data are presented in Table 1 and Figure 1. As before, erroneous responses and extremely long or short RTs (above 2000 ms or below 150 ms; 0.2% of the data) were removed. Training (trained competitor vs. no trained competitor) and test session (first vs. second vs. third) were included as fixed factors. By-subjects random slopes for test session were included, as they significantly improved the model fit. The factor of training contributed significantly to the model, $\chi^2(1)=9.33$, p=.002, as did test session, $\chi^2(2)=9.61$, p=.008. The interaction between training and test session was also significant, $\chi^2(2)=32.80$, p<.001. This interaction reflected the fact that while there was no training effect observed in the first test session, $\chi^2(1)=2.27$, p=.10, or the second session, $\chi^2(1)=0.30$, p=.59, there was a significant training effect in the third session, $\chi^2(1)=34.85$, p<.001.

Accuracy rates in the pause detection task are presented in Table 1 and were analysed as before. Training (trained competitor vs. no trained competitor) and test session (first vs. second vs. third) were included as fixed factors. No random slopes were included as they did not significantly improve the model fit. No significant main effects of training were found but a significant effect of test session did emerge, $\chi^2(2)=8.39$, p=.02. This reflected a significant difference in accuracy between sessions 1 and 2, z=2.40,p=.02, and between sessions 2 and 3, z=2.60,p=.01, no other contrasts were significant.

Free recall. One participant's data were lost in this task due to equipment failure. No random slopes were included in the model for free recall data. A significant main effect of test session was found, $\chi^2(2)=35.17$, p<.001. A comparison of recall rates across days showed no significant difference in recall rates between sessions 1 and 2, but recall in session 3 was

significantly higher than recall in session 1, z=5.37,p<.001, or session 2, z=4.84,p<.001 (Figure 2).

Old-new categorisation. Erroneous responses and extremely long or short RTs (above 3000 ms or below 500 ms; 0.2% of the data) were removed. Test session (first vs. second vs. third) was included as a fixed factor. By-subjects and by-items random slopes for the effect of session were retained. The main effect of session was significant, $\chi^2(2)=14.94$, p<.001. Pairwise comparisons of the three sessions showed a significant difference between sessions 2 and 3, $\chi^2(1)=12.87$, p<.001 and sessions 1 and 3, $\chi^2(1)=6.64$, p=.01, but no other contrasts were significant.

Accuracy data are presented in Figure 3. An ANOVA with test session as a within-participants factor showed no significant main effect of test session (p=.54).

Comparison across Experiments 1 and 3. Following the previous analysis, we combined pause detection data from the two experiments. Data were trimmed in the same manner as before. We entered training (trained competitor vs. no trained competitor), test session (first vs. second vs. third), and experiment (spoken vs. sung to familiar melody) as fixed factors. By-subjects random slopes for test session were retained. A significant three-way interaction was found, $\chi^2(2)=7.77$, p=.02, suggesting that the two experiments showed differences in lexical competition effects. To unpack this interaction, we analysed the three test sessions separately. In Session 1 we found no interaction between training and experiment. The main effect of training was not significant, demonstrating an absence of the lexical competition effect, but the main effect of experiment did reach significance, $\chi^2(1)=12.59$, p<.001, reflecting a general tendency for participants in Experiment 3 to make slower responses than participants in Experiment 1(Table 1). In Session 2 we found a trend-level interaction between training and experiment, $\chi^2(1)=3.24$, p=.07, consistent with the

significant lexical competition effect seen in Experiment 1 and the non-significant effect in Experiment 3. Both the main effect of training and the main effect of experiment reached significance, $\chi^2(1)=6.47$, p=.01 and $\chi^2(1)=11.40$, p<.001, respectively, with the former driven by the significant lexical competition effect seen in Experiment 1, and the latter again reflecting the tendency of participants in Experiment 3 to make slower pause detection responses across the conditions. Finally, in Session 3 we observed a significant interaction between training and experiment, $\chi^2(1)=3.82$, p=.05. As Figure 1 illustrates, this shows that in the final test session we observe a significantly stronger lexical competition effect when the novel words were learned sung to a familiar melody compared to when the words were learned in the spoken modality. In the pause detection error rates, the factor of experiment did not interact with any other factors.

Free recall data were analysed, with test session and experiment entered as fixed factors. Items-specific slopes were entered for the effect of experiment. We found no significant interaction, no main effect of experiment, and, again consistent with the individual analyses of Experiments 1 and 3, a significant main effect of test session, $\chi^2(2)=58.88$, p<.001.

Old-new categorisation RTs were analysed by entering experiment and test session as fixed factors. Subject and item-specific slopes for the effect of test session were retained. We found no significant interaction, but did find a main effect of experiment, $\chi^2(1)=12.82$, p<.001, and a significant main effect of test session, $\chi^2(2)=22.53$, p<.001. Accuracy data were analysed with an ANOVA as before. There was no interaction between the two factors, no main effect of test session, and a marginal main effect of experiment, $F(1,75)=3.27$,p=.075.

We found no significant difference between the two experiments in training performance (p=.41). As before, this suggests that the novel words in both experiments were equally intelligible and that participants attended to the training equally well.

Discussion

The pause detection task replicated findings from Experiments 1 and 2 in that we saw no lexical competition effects in the first session, and observed a significant effect in the last session. However, in earlier experiments the competition effect emerged in the second session, one day after training: in the current experiment the effect was not significant until the third and final session, suggesting that training in the sung modality using a familiar melody may have led to a delayed integration of the novel words in the mental lexicon. It is important to point out that the interaction between the competition effect and experiment in session 2 was only marginally significant (p=.07), therefore this delayed competition effect should be interpreted with a degree of caution. Another difference between the experiments was seen in session 3. Here the lexical competition effect was significantly larger in Experiment 3 than in Experiment 1, as manifested by the significant interaction between the competition effect and experiment.

We found no difference between the spoken and sung modalities in the free recall task, even when using a familiar melody during training. Like in the first two experiments, recall rates increased in the final test session, again probably due to increased number of exposures to novel words during testing.

In the old-new categorisation task we found lower accuracy rates in the current experiment than in Experiment 1 (although the effect here was marginal), replicating to a

lesser extent the modality-related accuracy difference seen in Experiment 2. However, in the present experiment the accuracy difference was accompanied by an RT difference; participants were significantly slower to make the decisions in Experiment 3 than in Experiment 1. The free recall and the old-new categorisation tasks therefore both suggest that explicit memory for novel words does not benefit from musical presentation.

General Discussion

We reported three experiments examining the impact of musical presentation on word learning and the integration of new words in the mental lexicon. In Experiment 1 we trained participants on novel words in the spoken modality. We found no evidence of these novel words being integrated in the mental lexicon immediately after training, as measured by the lexical competition effect. However, evidence of integration emerged one day after training and remained robust one week later. This finding successfully replicates earlier work using this learning paradigm (e.g., Tamminen & Gaskell, 2008). In Experiment 2 we trained the same novel words in the sung modality using an unfamiliar isochronous melody. This addition did not change the emergence of lexical competition effects from the trajectory seen with spoken training; there was no difference in the magnitude or the time course of the effect compared to the spoken modality training. In Experiment 3 we did not observe lexical competition effects in the sung modality until the last test session, a week after training. However, when the lexical competition effect did emerge in this condition, it was significantly larger than in the spoken training experiment.

We had hypothesised that if musical presentation engages a broader neural network beyond the MTL during training compared to speech, we might have observed accelerated lexical integration, with integration effects emerging immediately after training. While this

was not the case, we did see stronger lexical integration effects on day 8 in Experiment 3 compared to those observed with spoken presentation (Experiment 1). This finding suggests that presentation of novel words in combination with familiar music may have resulted in stronger lexical representations, or representations that are more strongly connected to phonologically overlapping competitors. More research is needed to understand precisely why familiar musical presentation, and which aspects of musical features, might have this effect. One possibility arises from studies suggesting that music increases neocortical plasticity at the time of encoding, for example by increasing the coherence of oscillatory brain processes (Thaut, et al., 2014; Peterson & Thaut, 2007). Another possibility is that familiar musical sound can facilitate the encoding of multicomponent (i.e. musical and verbal) feature bound auditory representations of novel words, a form of additive binding (de Vignemont, 2015), with benefits for redintegration processes in memory (Williamson, Baddeley & Hitch, 2010).

We also note with interest the trend-level finding of a delayed lexical integration effect in Experiment 3. Delayed lexical competition effects in a word learning study have been previously reported by Bakker et al. (2014). When participants were trained on the novel words in the visual modality (i.e. participants saw the words during training but never heard them) and tested in the auditory modality using the pause detection task, lexical competition effects were observed only one week after training began. When there was no change in modality from training to test (i.e. both training and test occurred either in the visual or auditory modality), the typical 24-hour delay in lexical integration was seen. The authors suggest the delayed competition effect is due to cross-modal lexical representations taking more time to emerge than same-modality representations. Our results suggest that this might be the case not only in the visual-auditory domain but also in the sung-spoken domain

(at least when familiar melodies are used, perhaps because they are more readily encoded alongside the words).

Given that music has been shown to benefit free recall performance in populations with and without memory impairments, as outlined in the Introduction, we hypothesised that our sung training modality might result in better free recall performance than the spoken modality. However, free recall rates were unaffected by training modality. This discrepancy may arise from differences in the musical stimuli across paradigms. One general mechanism of music-related task improvement is an increase in psychophysiological arousal and mood (Cassidy & MacDonald, 2007; Furnham & Strbac, 2002; Thompson, Schellenberg, & Husain, 2001). It is possible that our sung stimuli did not provide a baseline level of increased stimulation compared to spoken stimuli that previous studies achieved. However, this remains speculation as psychophysiological arousal is not commonly monitored in memory studies, therefore it is not clear that changes are a pre-requisite for improved performance. Future studies of word learning and music would be advised to monitor arousal and mood during the task in order to investigate this hypothesis further.

The discrepancy between our recall findings and those in the literature may also be explained by other basic paradigm differences. The most popular approach is to teach participants new lyrics or verse with either spoken or sung presentation (McElhinney & Annett, 1996; Wallace, 1994; Kilgour, Jakobson, & Cuddy, 2000; Racette & Peretz, 2007) where the verbal stimuli contain known words that are linked in sentence form. These additional phonological and semantic cues may derive support from musical structures, most notably the existence of rhythmic boundaries and phrase patterns to guide redintegration of verbal materials from long-term memory (Purnell-Webb & Speelman, 2008; Cason & Schön, 2012). Our melodies would have provided limited support in both these regards since the words were novel and had the same syllable structure, and the lists had no long-term

dependencies or patterns. Furthermore, the melodies were isochronous and sung with minimal stress, meaning that participants were not able to utilise rhythm and time-related phrasing cues that would be present in typical singing. We also acknowledge that the way participants were trained was quite unusual compared to the way people learn new words in natural settings. Recent studies using the same test tasks but a more ecologically valid training method however have shown similar levels of learning and consolidation effects as we show here (Henderson, Devine, Weighall, & Gaskell, 2015) suggesting that our training method did not impede learning compared to more natural methods.

Accuracy rates in the old-new categorisation task were significantly lower when the words had been learned in the sung modality. In Experiment 3, when participants were made familiar with the sung melody, the lower performance in old-new categorisation accuracy rates extended to reaction times, with RTs being significantly slower in Experiment 3 than in Experiment 1. While these data clearly show a lack of a learning advantage in the sung modality, it is important to note that they do not indicate a learning *disadvantage*. These accuracy and RT results are to be expected due to the fact that there was a modality mismatch between training (sung) and the test stimuli (spoken) rather than a specific negative impact of musical presentation. This is supported by a similar modality mismatch effect in the same task reported by Brown and Gaskell (2014) who manipulated the voice in which participants learned novel words and were later tested. When the voice was different at training and at test (female vs. male), old-new categorisation accuracy rates were significantly lower than when the voice matched. Brown and Gaskell (2014) argued that this shows that voice-specific details are encoded and stored in lexical memory, and that this information is retained for at least a week. Our experiments suggest that information about the sung vs. spoken modality is retained and treated much in the same way as speaker identity. If on the other hand the sung modality improved recognition memory over the spoken modality, the modality mismatch

effect might have been overcome in Experiments 2 and 3, or if the melody familiarity manipulation improved recognition memory, the modality mismatch effect should have been attenuated in Experiment 3 compared to Experiment 2. Neither of these hypotheses were supported by the data, and it therefore appears that recognition memory, like free recall, was not affected by the sung training modality.

Overall, musical presentation seems to have a complex role to play in verbal learning. Previous studies demonstrated a role for familiar music in improving episodic memory for sung lyrics and phrases, perhaps because it provides multiple cues for the redintegration of sequences from long-term memory. The present work builds on this finding by showing that hearing unrelated completely novel words (i.e. new vocabulary) sung to familiar isochronous melodies does not have any positive impact on episodic memory, but may support further stages of learning, by virtue of enhancing lexical integration over the course of one week. Future controlled studies are needed to confirm precisely which dimensions of music best facilitate new learning, and whether dimensions which were not manipulated here, such as rhythm and phrasing, as well as lifelong familiarity with melodies, would have an impact on episodic memory tasks like free recall and recognition memory.

Acknowledgements

This research was funded by a British Academy small grant to JT and VJW, and by an Experimental Psychology Society small grant to JT. JT was also supported by a British Academy Postdoctoral Research Fellowship, and VJW by a Vice Chancellor's Fellowship from the University of Sheffield. We thank Pierre Gagnepain and Matt Davis for sharing their novel word stimuli, Marcus Pearce for sharing his hymnal database and computational model, and Sagar Jilka for recording early versions of the melodies.

References

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language, 59,* 390–412.

Baayen, R. H., Piepenbrock, R., & van Rijn, H. (1995). The CELEX Lexical Database (CD-ROM). Linguistic Data Consortium, University of Pennsylvania, Philadelphia, PA.

Bakker, I., Takashima, A., van Hell, J. G., Janzen, G., & McQueen, J. M. (2014). Competition from unseen or unheard novel words: Lexical consolidation across modalities. *Journal of Memory and Language, 73,* 116-130.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language, 68,* 255–278.

Baur, B., Uttner, I., Ilmberger, J., Fesl, G., & Mai, N. (2000). Music memory provides access to verbal knowledge in a patient with global amnesia. *Neurocase, 6,* 415-421.

Brown, H., & Gaskell, M. G. (2014). The time-course of talker-specificity and lexical competition effects during word learning. *Language, Cognition, and Neuroscience, 29,* 1163-1179.

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,* 245-260.

Cason, N., & Schön, D. (2012). Rhythmic priming enhances the phonological processing of speech. *Neuropsychologia, 50,* 2652–2658.

Cassidy, G., & MacDonald, R. (2007). The effect of background music and background noise on the task performance of introverts and extraverts. *Psychology of Music*, 35, 517–537.

Davis, M. H., Di Betta, A. M., Macdonald, M. J. E., & Gaskell, M. G. (2009). Learning and consolidation of novel spoken words. *Journal of Cognitive Neuroscience, 21,* 803-820.

Davis, M. H., & Gaskell, M. G. (2009). A complementary systems account of word learning: Neural and behavioural evidence. *Philosophical Transactions of the Royal Society, 364,* 3773-3800.

de Vignemont, F. (2015). Multimodal unity and multimodal binding. In C. Mole and D. Bennett (eds.) Sensory integration and the unity of consciousness. MIT Press, forthcoming.

Dickson, D., & Grant, L. (2003). Physics karaoke: Why not? *Physics Education, 38,* 320-323.

Dumay, N., & Gaskell, M. G. (2007). Sleep-associated changes in the mental representation of spoken words. *Psychological Science, 18,* 35-39.

Dumay, N., & Gaskell, M. G. (2012). Overnight lexical consolidation revealed by speech segmentation. *Cognition, 123,* 119–132.

Forster, K. I., & Forster, J. C. (2003). DMDX: A Windows display program with millisecond accuracy. *Behavior Research Methods, Instruments, & Computers, 35,* 116-124.

Furnham, A., & Strbac, L. (2002). Music is as distracting as noise: The differential distraction of background music and noise on the cognitive test performance of introverts and extraverts. *Ergonomics, 45,* 203–217.

Gaskell, M. G., & Dumay, N. (2003). Lexical competition and the acquisition of novel words. *Cognition, 89,* 105-132.

Good, A., Russo, F. A., & Sullivan, J. (2015). The efficacy of singing in foreign-language learning. *Psychology of Music, 43,* 627-640.

Haslam, C., & Cook, M. (2002). Striking a chord with amnesic patients: Evidence that song facilitates memory. *Neurocase, 8*, 453-465.

Henderson, L., Devine, K. Weighall, A. Gaskell, M. G. (2015). When the daffodat flew to the intergalactic zoo: Off-line consolidation is critical for word learning from stories. *Developmental Psychology, 51,* 406-417.

Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language, 59,* 434–446.

Kang, H.-J., & Williamson, V. J. (2014). Background music can facilitate second language learning. *Psychology of Music*, *42,* 728-747.

Kilgour, A. R., Jakobson, L. S., & Cuddy, L. L. (2000). Music training and rate of presentation as mediators of text and song recall. *Memory & Cognition, 28,* 700-710.

Koelsch, S. (2011). Toward a neural basis of music perception - a review and updated model. *Frontiers in Psychology, 2,* 1-20.

Marslen-Wilson, W. (1987). Functional parallelism in spoken word recognition. *Cognition, 25,* 71-102.

Mattys, S. L., & Clark, J. H. (2002). Lexical activity in speech processing: evidence from pause detection. *Journal of Memory and Language, 47,* 343-359.

McClelland, J. L., McNaughton, B. L., & O'Reilly, R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychological Review, 102,* 419-457.

McElhinney, M., & Annett, J. M. (1996). Pattern of efficacy of a musical mnemonic on recall of familiar words over several presentations. *Perceptual & Motor Skills, 82,* 395-400.

Medina, S. L. (1993). The effect of music on second language vocabulary acquisition. *National Network for Early Language Learning, 6,* 1–8.

Moussard, A., Bigand, E., Belleville, S., & Peretz, I. (2012). Music as an aid to learn new verbal information in Alzheimer's disease. *Music Perception, 29,* 521-531.

Moussard, A., Bigand, E., Belleville, S., & Peretz, I. (2014). Learning sung lyrics aids retention in normal ageing and Alzheimer's disease. *Neuropsychological Rehabilitation, 24,* 894-917.

Mueller, K., Fritz, T., Mildner, T., Richter, M., Schulze, K., Lepsien, J., Schroeter, M. L, Möller, H. E. (2015). Investigating the dynamics of the brain response to music: A central role of the ventral striatum/nucleus accumbens. *NeuroImage, 116*, 68-79.

Nicholson, S., Knight, G. H., Dykes Bower, J. (1950). *Ancient and modern revised*. William Clowes and Sons, Suffolk, UK.

Omigie, D., Pearce, M. T., Williamson, V. J., & Stewart, L. (2013). Electrophysiological correlates of melodic processing in congenital amusia, *Neuropsychologia, 51,* 1749–1762.

Palisson, J., Roussel-Baclet, C., Maillet, D., Belin, C., Ankri, J., & Narme P. (2015). Music enhances verbal episodic memory in Alzheimer's disease. *Journal of Clinical Experimental Neuropsychology, 8,* 1-15.

Pearce, M. T. (2005). The construction and evaluation of statistical models of melodic structure in music perception and composition. Doctoral Dissertation, Department of Computing, City University, London, UK.

Pearce, M. T., & Wiggins, G. A. (2006). Expectation in melody: the influence of context and learning. *Music Perception, 23,* 377–405.

Pearce, M. T., Ruiz, M. H., Kapasi, S., Wiggins, G. A., & Bhattacharya, J. (2010). Unsupervised statistical learning underpins computational, behavioural, and neural manifestations of musical expectation. *NeuroImage, 50,* 302–313.

Peretz, I., & Zatorre, R. J. (2005). Brain organization for music processing. *Annual Review of Psychology, 56*, 89-114.

Peterson, D. A., & Thaut, M. H. (2007). Music increases frontal EEG coherence during verbal learning. *Neuroscience Letters, 412,* 217-221.

Purnell-Webb, P., & Speelman, C. P. (2008). Effects of music on memory for text. *Perceptual & Motor Skills, 106,* 927-957.

Racette, A., & Peretz, I. (2007). Learning lyrics: To sing or not to sing? *Memory & Cognition, 35,* 242-253.

Simmons-Stern, N. R., Budson, A. E., & Ally, B. A. (2010). Music as a memory enhancer in patients with Alzheimer's disease. *Neuropsychologia, 48,* 3164-3167.

Simmons-Stern, N. R., Deason, R. G., Brandler, B. J., Frustace, B. S., O'Connor, M. K., Ally, B. A., & Budson, A. E. (2012). Music-based memory enhancement in Alzheimer's disease: Promise and limitations. *Neuropsychologia, 50,* 3295-3303.

Tamminen, J., Davis, M. H., & Rastle, K. (2015). From specific examples to general knowledge in language learning. *Cognitive Psychology, 79,* 1-39.

Tamminen, J., & Gaskell, M. G. (2008). Newly learned spoken words show long-term lexical competition effects. *Quarterly Journal of Experimental Psychology, 61,* 361-371.

Tamminen, J., Payne, J. D., Stickgold, R., Wamsley, E. J., & Gaskell, M. G. (2010). Sleep spindle activity is associated with the integration of new memories and existing knowledge. *Journal of Neuroscience, 30,* 14356–14360.

Thaut, M. H., Peterson, D. A., McIntosh, G. C., & Hoemberg, V. (2014). Music mnemonics aid verbal memory and induce learning-related brain plasticity in multiple sclerosis. *Frontiers in Human Neuroscience, 8,* 1-10.

Thompson, W. F., Schellenberg, E. G., & Husain, G. (2001). Arousal, mood, and the Mozart effect. *Psychological Science, 12,* 248–251

Wallace, W. T. (1994). Memory for music: Effect of melody on recall of text. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 20,* 1471-1485.

Williamson, V. J., Baddeley, A. D., & Hitch, G. J. (2010). Musicians' and nonmusicians' short-term memory for verbal and musical sequences: Comparing phonological similarity and pitch proximity. *Memory and Cognition, 38,* 163-175.

Table 1. Means of accurate pause detection RTs (in ms ± standard error) and error rates to base words (e.g., *dolphin*). Percentage of errors in parentheses. Trained base words are base words for which a new competitor was learned, no new competitor was included in the training for untrained base words.

		Session 1	Session 2	Session 3
Experiment 1	Trained	578±18	581±18	593±17
		(5.0±0.9%)	(4.1±0.8%)	(5.0±0.9%)
	Untrained	578±16	558±16	571±17
		(5.0±1.0%)	(5.3±1.1%)	(4.5±1.1%)
Experiment 2	Trained	620±17	608±21	690±31
		(4.0±0.7%)	(3.9±0.8%)	(6.3±1.0%)
	Untrained	613±17	593±21	645±26
		(4.0±0.8%)	(4.8±0.9%)	(4.6±0.8%)
Experiment 3	Trained	657±16	655±18	727±21
		(5.3±1.1%)	(3.9±0.7%)	(6.0±1.0%)
	Untrained	672±18	652±18	674±16
		(4.8±1.1%)	(3.4±0.6%)	(4.4±0.9%)

Figure Captions

Figure 1. Magnitude of the lexical competition effect in all three test sessions in all three experiments. The competition effect is calculated by deducting pause detection RTs to base words with no newly learned competitor from RTs to base words with a new competitor. Therefore positive numbers indicate that the newly learned words are engaging in lexical competition. Error bars indicate standard error.

Figure 2. Proportion of newly learned words recalled in the free recall task in all three test sessions across all three experiments. Error bars indicate standard error.

Figure 3. Accuracy rates (in d') and RTs (in ms) in the old-new categorisation task in all three test sessions across all three experiments. The bar graph shows accuracy rates while the line graph shows RTs. Error bars indicate standard error.

Figure 4. An example of a sung melody as heard by participants in Experiments 2 and 3.

Figure 1.

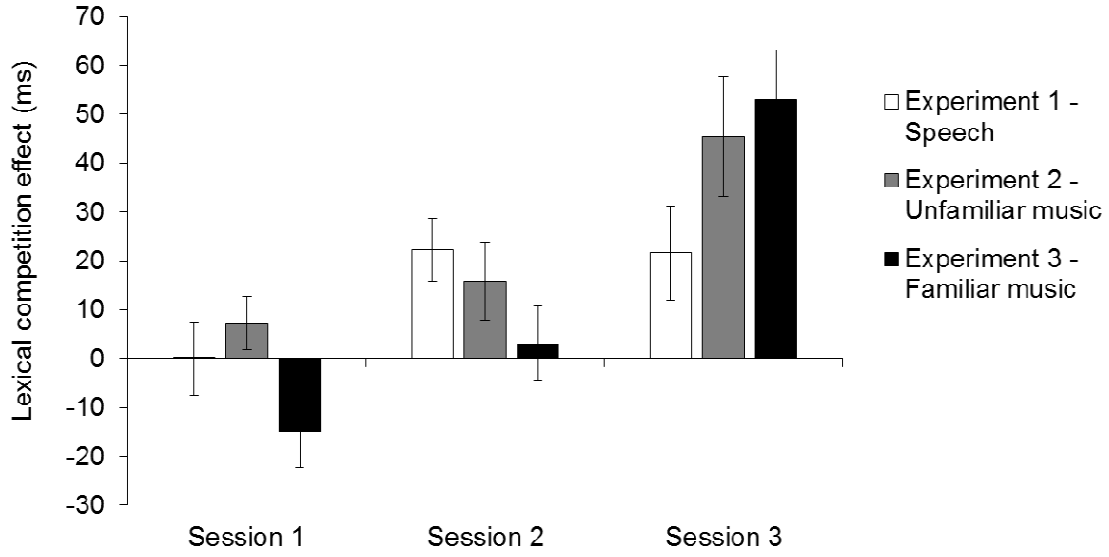

70
60
50
40
30
20
10
0
-10
-20
-30
Lexical competition effect (ms)
Session 1
Session 2
Session 3
Experiment 1 - Speech
Experiment 2 - Unfamiliar music
Experiment 3 - Familiar music

Figure 2.

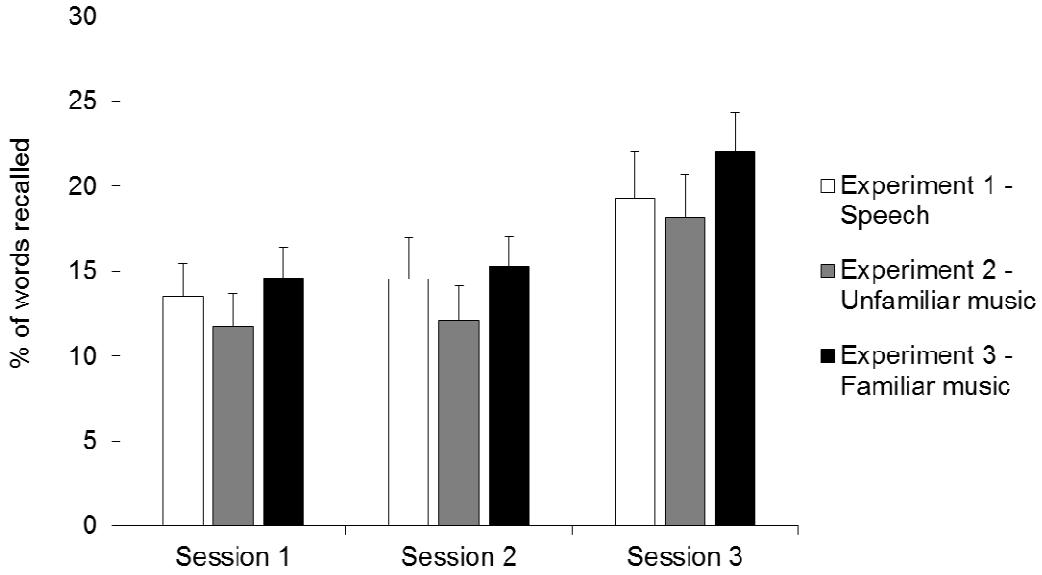

30
25
20
15
10
5
0
% of words recalled
Session 1
Session 2
Session 3
Experiment 1 - Speech
Experiment 2 - Unfamiliar music
Experiment 3 - Familiar music

Figure 3.

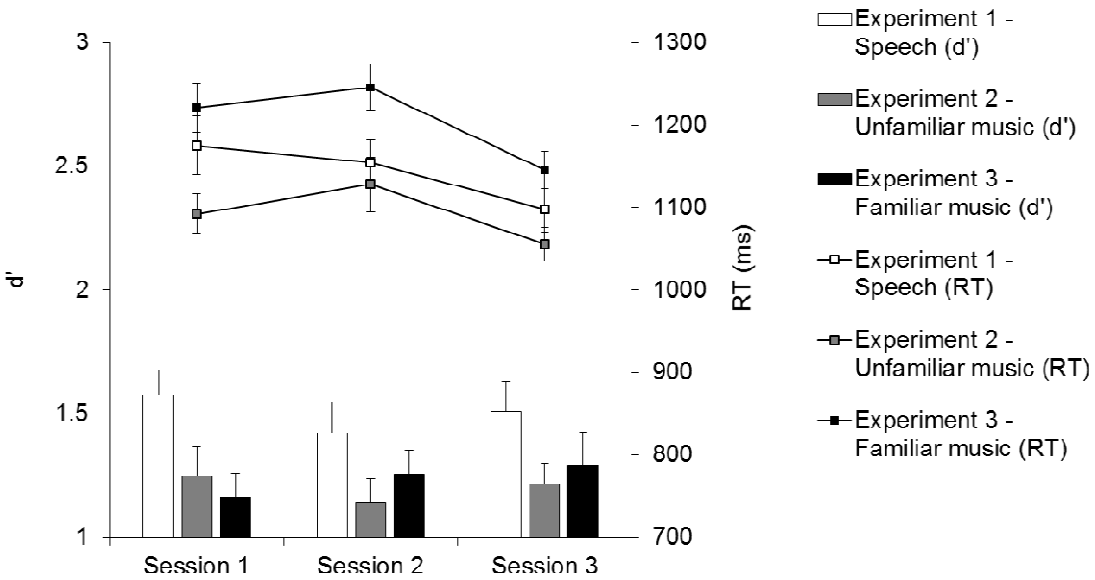
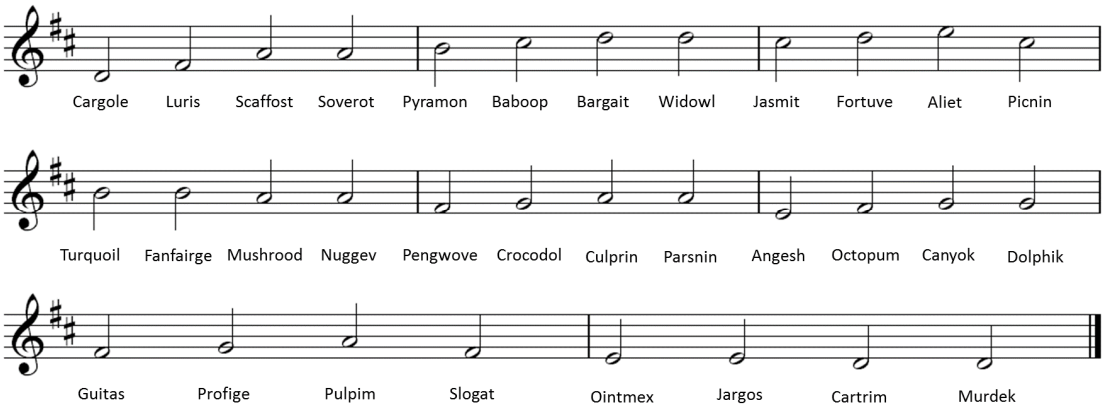

3
2.5
2
1.5
1
d'
1300
1200
1100
1000
900
800
700
RT (ms)
Session 1
Session 2
Session 3
Experiment 1 - Speech (d')
Experiment 2 - Unfamiliar music (d')
Experiment 3 - Familiar music (d')
Experiment 1 - Speech (RT)
Experiment 2 - Unfamiliar music (RT)
Experiment 3 - Familiar music (RT)

Figure 4.


Cargole Luris Scaffost Soverot Pyramon Baboop Bargait Widowl Jasmit Fortuve Aliet Picnin
Turquoil Fanfairge Mushrood Nuggev Pengwove Crocodol Culprin Parsnin Angesh Octopum Canyok Dolphik
Guitas Profige Pulpim Slogat Ointmex Jargos Cartrim Murdek