Behav Res (2018) 50:684–702
DOI 10.3758/s13428-017-0894-6

The MUSOS (MUsic SOftware System) Toolkit: A computer-based, open source application for testing memory for melodies

M. Rainsford[1] · **M. A. Palmer**[1] · **G. Paine**[2]

Published online: 21 April 2017

Abstract Despite numerous innovative studies, rates of replication in the field of music psychology are extremely low (Frieler et al., 2013). Two key methodological challenges affecting researchers wishing to administer and reproduce studies in music cognition are the difficulty of measuring musical responses, particularly when conducting free-recall studies, and access to a reliable set of novel stimuli unrestricted by copyright or licensing issues. In this article, we propose a solution for these challenges in computer-based administration. We present a computer-based application for testing memory for melodies. Created using the software Max/MSP (Cycling '74, 2014a), the MUSOS (Music Software System) Toolkit uses a simple modular framework configurable for testing common paradigms such as recall, old–new recognition, and stem completion. The program is accompanied by a stimulus set of 156 novel, copyright-free melodies, in audio and Max/MSP file formats. Two pilot tests were conducted to establish the properties of the accompanying stimulus set that are relevant to music cognition and general memory research. By using this software, a researcher without specialist musical training may administer and accurately measure responses from common paradigms used in the study of memory for music.

Keywords Music cognition · Software · Replication · Memory · Recognition

✉ M. Rainsford
miriam.rainsford@utas.edu.au

[1] Psychology, School of Medicine, University of Tasmania, Sandy Bay, Tasmania, Australia

[2] School of Arts, Media and Engineering, Arizona State University, Tempe, AZ, USA

Music psychology is an emerging field of research that has contributed numerous theoretical models to the literature describing the ways in which musical elements such as pitch, melody, and harmony are perceived, processed, and remembered (Deutsch, 1982; Krumhansl, 1991; Snyder, 2000, 2009). Insights gained from research into the cognition of music have also contributed to our understanding of general cognitive processes. The study of memory for musical melodies has yielded insights into the way in which auditory material is perceived and encoded, leading to an improved understanding of working memory processes (Berz, 1995; Williamson, Baddeley, & Hitch, 2010), and the identification of differences between verbal and musical semantic memory (Schulkind, 2004). However, despite considerable growth in the music psychology literature over the last 30 years, independent evidence confirming the reproducibility of findings is lacking (Frieler et al., 2013). As in general psychology (Open Science Collaboration, 2015), there is a pressing need to facilitate replication studies in music cognition. According to a recent review by Frieler and colleagues the percentage of exact replication studies and meta-analyses published in four major music psychology journals is around 1%, with only ten meta-analyses and 18 replication studies identified overall. In music cognition, the difficulty of developing and administering accurate measures of participant response further compounds the task of replicating previous findings. Considerable advances have been made in the measurement and understanding of participant responses through computer-based analysis (Müllensiefen & Wiggins, 2011). In this article, we present a computer-based toolkit designed to help researchers overcome two key problems faced when designing and replicating music cognition studies: measurement of recall responses and the availability of novel stimuli.

The first problem concerns measuring and interpreting participants' responses in studies of music and memory. Fewer

studies have been undertaken of musical recall than recognition (Müllensiefen & Wiggins, 2011), as challenges are presented in recording and interpreting an accurate response from untrained musicians (Sloboda & Parker, 1985). Where test administration involves musical performance at a keyboard, or the interpretation of sung responses from a participant (e.g., Bailes, 2010; Warker & Halpern, 2005), a researcher with skilled musical training is required, further limiting the replicability of studies.

Computer-based data analysis has facilitated improvements in the interpretation of musical data, allowing participant responses to be interpreted objectively and with greater accuracy (Müllensiefen & Wiggins, 2011). A computer-based method for testing paradigms of musical recognition and recall would ensure that a participant's true response is being measured, while reducing reliance on trained musicians as researchers.

We present a computer-based method for testing memory for musical melodies. Designed in Max/MSP 6.1 (Cycling '74, 2014a), the MUSOS (MUsic SOftware System) Toolkit is compatible with computers running Windows XP and above, and Mac OS X 10.5 and above. The application consists of a framework housing several modules that may be configured to administer standard paradigms used in memory research including recall, explicit recognition, and implicit memory studies including stem completion. The program is open source, released under the Gnu General Public License (GPL) 3.0 (Free Software Foundation, 2007), with documentation provided on configuring the modules provided to create tests of different types and stimulus length. The toolkit, including all source files, documentation, and sample data, is available for download at http://www.soundinmind.net/MUSOS/MUSOS.zip. An experienced Max/MSP programmer is welcome to download and customize the program according to their needs.

The second problem faced by researchers in music cognition concerns the availability of novel musical stimuli. In general, studies of musical memory have used databases of folk songs (e.g., Bailes, 2010; Schmuckler, 1997), which are out of copyright but present the possibility that an unknown folk melody may trigger memory for other, similar folk songs (see Sloboda & Parker, 1985, p. 159). Alternatively, databases of popular songs already known to the participant have been used to test online recognition and absolute pitch memory (e.g., Jakubowski & Müllensiefen, 2013; Levitin, 1994; Schulkind, 2004). Although database sources are commonly used as an accessible means to make stimuli available, the researcher may wish to control the degree to which participants are exposed to the melodies, rather than relying on exposure via popular media or other external sources. A novel set of 156 copyright-free melodic stimuli is therefore provided with the MUSOS Toolkit, comprising a set of 78 eight-note and a set of 78 sixteen-note melodies. All melodies are composed on a non-Western modal scale, so as to reduce the possibility that sources outside of the laboratory are triggered in memory. The stimuli were analyzed using the application FANTASTIC (Müllensiefen, 2009a) for properties important in the study of music cognition, including pitch, intervallic, and contour features. The stimuli were also rated by a group of pilot testers for values of distinctiveness and valence, variables that have been found to be associated with improved memory for musical items (Bailes, 2010; Stalinski & Schellenberg, 2013). The stimulus set is released under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International license (Creative Commons, 2013); thus, no copyright issues are presented for researchers who wish to use these melodies in testing or to reproduce examples in a journal article. The stimuli are supplied as both Max/MSP jit.cellblock text files and in wav file format, so that they may be imported into an existing software framework if preferred. The program may also be configured so that researchers may enter and save their own stimulus sets.

In this article, we first describe the rationale and design of the MUSOS software application and the tests for which it may be configured. We further describe the method used in constructing and testing the accompanying stimulus set. We present results from two pilot tests, the first conducted to obtain values describing features of the accompanying stimulus set important to studies of music and memory. The second pilot test was conducted to establish a subset of stimuli from the provided collection that were designed to be either very difficult or very easy to remember. The data obtained in pilot testing thus enables researchers to use MUSOS and its accompanying stimulus set "out of the box" to set up studies.

Understanding and measuring memory for musical stimuli

In developing a stimulus set to accompany a toolkit for studies in music and memory, it is important to consider the ways in which musical information is perceived, stored and retrieved, and the factors that influence successful retrieval. We provide below a brief introduction to auditory memory and processing of the melodic features for which we provide measurement in the accompanying stimulus set, however, for a comprehensive introduction to the topic of music and memory we recommend the seminal works of Deutsch (1982), Snyder (2000), and the *Oxford Handbook* (Hallam, Cross, & Thaut, 2009).

The present cognitive model of auditory memory ingrates Sperling (1960) and Darwin, Turvey, and Crowder's (1972) concept of a brief sensory (echoic) memory store, with Baddeley and Hitch's (1974; Baddeley, 2012) model describing the transfer of incoming perceptual information from sensory memory both to processing and rehearsal in working memory and to storage and retrieval from long-term memory.

As for other domains, long-term memory may be implicit, without conscious awareness, or explicit (Schacter, 1987). Although memory for musical structures, like language, is stored in semantic memory, episodic memory is also involved in remembering experiences of music (Snyder, 2000).

Incoming auditory information from the environment is initially perceived by the nerve cells of the ear as a series of impulses representing frequencies and amplitudes. Auditory information is then stored in echoic memory as a very brief sensory image, lasting only a few seconds (Darwin et al., 1972; Snyder, 2009). At this stage, features occurring simultaneously or close together are extracted from the incoming information stream by higher level neurons and bound into units so that they may be perceived *categorically* as separate pitches, and interval relationships between pitches (Aruffo, Goldstone, & Earn, 2014; Snyder, 2000). Categorical perception of pitch, interval distances, and basic rhythmic features is a *bottom-up* process in which the information stream is grouped by the nervous system and perceived as events (Dowling, 1982; Snyder, 2000). Larger level groupings occur as information is passed from echoic to working memory; events occurring sequentially are bound together and perceived as rhythmic patterns, or brief melodic phrases. The process of feature extraction and categorical perception may at the same time trigger recognition, through activation of previously stored experiences in long-term memory (Snyder, 2000).

Working memory is limited in capacity, and can store approximately seven (plus or minus two) unique items (Miller, 1956). Information in working memory must be *rehearsed* in order to be stored in long-term memory (Baddeley & Hitch, 1974). The amount of information being manipulated in working memory may be increased through grouping or "chunking" into repeated patterns. In music, this may involve repetition of sequences of notes and rhythmic patterns to build a complete phrase; the length of a musical phrase is often designed to be approximately the same duration as the capacity of working memory, on average around 4–8 s. Larger-scale groupings of phrases into formal musical structures are understood and stored in long-term memory (Snyder, 2000, 2009).

Working memory is currently understood to have at least four components, these being a *central executive*, which coordinates operations on information held in three buffers used to process different types of sensory material, the *visuospatial sketchpad*, the *phonological loop*, involved in the rehearsal and storage of verbal material, and the *episodic buffer*, which stores brief episodic experiences (Baddeley, 2012; Baddeley & Hitch, 1974). Verbal and auditory information are proposed to share use of the phonological loop, however, recent evidence also supports a separate store for musical pitch, as a *tonal loop* involved in the rehearsal of tonal information (Schulze, Zysset, Müller, Friederici, & Kölsch, 2010; Williamson et al., 2010). Music, however, does not involve just a single store, but is a multisensory experience integrating auditory, episodic and visual processing (Williamson et al., 2010).

Grouping of information into chunks may occur either through bottom-up processing of information at the psychophysical level (Dowling, 1982), or through *top-down*, schema-driven processing (Snyder, 2009), in which previous experiences define a set of schemata or higher-order abstractions, through which a listener may understand, recognize, or make predictions about a piece of music (Deutsch, 1999; Krumhansl, 1991). These may include information on pitch chroma hierarchies, tonality, contour, and rhythmic patterns, as well as relationships between these features (Snyder, 2009).

The processing of pitch material has most notably been investigated by Deutsch (1970, 1972, 1973, 1974, 1975), who proposed that neural pathways involved in the processing of musical pitch are organized hierarchically in a similar way to the perception of letters and words. Most musicians, unless they possess absolute pitch, recognize a melody from its *intervals*, or the distance in semitones between consecutive notes. Deutsch (1969) proposed that a lower-level neural system dedicated to the recognition of musical intervals in turn activates a higher-level organization of neurons based around the musical scale, thus explaining the recognition and storage of melodies in terms of their intervallic structure and relationship to musical scale.

Although basic pitch and interval distance perception involves bottom-up processes at the psychophysical level (Deutsch, 1999; Dowling, 1982), Deutsch (1972) obtained evidence that, similar to verbal information, interval perception is also informed by top-down processes. When a well-known folk melody was presented to participants with the octave placement of its notes randomly varied, or with pitch information removed, listeners were unable to recognize the melody. However, when the name of the tune was provided, listeners were able to follow the melody, by matching the perceived tones against their expectations of intervallic relationships (Deutsch, 1999).

Krumhansl (1991); Krumhansl & Kessler, 1982) further demonstrated schema-based processing of hierarchical relationships between the notes of the scale, or pitch chroma, as certain notes are perceived as closer or more distant to the root note of the scale or *tonic*. Schemata defining these relationships are acquired implicitly from music-listening, and vary according to the listener's exposure to cultural musical traditions (Stevens & Byron, 2009). In Western music, notes of the scale close to the tonal center, and intervals based on close relationships to the tonic (e.g., perfect fourth and fifth) are more predictable (Bailes, 2010). Following from this, melodies that are more *tonal*—that is, whose content is built around such strong relationships to the tonic—are more expectable, and thus better remembered (Deutsch, 1980; Krumhansl, 1991; Schmuckler, 1997; Vuvan, Podolak, & Schmuckler, 2014). Melodies containing such schema-congruent, or in musical-theoretical terms, tonal events are also perceived as more pleasurable (Huron, 2006). At the same time, Vuvan and

colleagues found a U-shaped relationship to expectancy, such that a distinctiveness effect (Schacter & Wiseman, 2006) also occurs in memory for melodies. Both highly expected, and highly unexpected notes in relationship to the tonic facilitate improved memory.

In addition to scale and tonal relationships, the contour, or rise and fall of a melody, plays an important role in melodic recognition. White (1960) and later Dowling (1978; Dowling & Fujitani, 1971) demonstrated that melodies may be recognized by their contour, even when individual notes are distorted. Melodies are also easier to discriminate when their contours are different, but discrimination between a standard and comparison melody is more difficult when a melody is subject to *tonal transposition*, where the contour is retained but the notes of the melody are shifted upward along the same scale, altering its intervals slightly. From this evidence, Dowling (1978) proposed that musical contour is processed and stored independently from memory for pitch and interval sizes. Where the tonal context of a melody is ambiguous (e.g., in tonal transpositions, or atonal melodies) the listener relies upon contour to recognize melodies (Dowling, 1982). The ability to discriminate contour develops in infancy, along with the ability to reproduce pitch and understand basic rhythmic groupings, whereas discrimination of intervals and schema-based processing of tonality begins later in childhood, developing toward adulthood (Dowling, 1982).

Halpern (1984) discovered a similar hierarchy in the priority to which non-musicians and musicians process scale, contour and rhythmic content of melodies. When encountering novel music, melodies are initially discriminated on the basis of their rhythmic content, followed by contour. For non-musicians, mode (whether the melody is written on a major or minor scale) is the least salient element, further demonstrating the importance of contour in melodic recognition, although mode was found to have greater importance to trained musicians.

It is therefore important that a researcher wishing to study music and memory has access to information describing the pitch and intervallic relationships, tonality, and contour of the stimuli to be used, to determine which stimuli are likely to be perceived and remembered with greater or lesser ease. Various computational methods have been developed to measure these factors in melodies. In this study, we used Müllensiefen's (2009a) application FANTASTIC to measure the stimuli provided with the MUSOS Toolkit. This software is capable of producing descriptive statistics and measures of entropy describing the uncertainty or predictability of pitch content (tone chroma), intervallic content, and the degree to which the melody accords to major or minor scale tonality. Contour is described using Huron's (1996) eight classifications, and Steinbeck's (1982) step contour and interpolation contour methods.

Paradigms used in the study of memory for music

In selecting paradigms for inclusion in a toolkit designed to facilitate studies of music and memory, one must consider not only the applicability of the paradigms to be included and their relevance to the literature, but also the architecture and usability of the program. Scientific software is frequently developed by specialist end-users, restricting further development to the laboratory where the software was developed (Macaulay et al., 2009). Similarly, reliance on specialist knowledge can potentially restrict studies of music psychology to a single laboratory or group of researchers. If we aim to create tools that make administration and retesting of studies easier for a non-specialist researcher, then the architecture of that software must be logically designed to facilitate ease of use (John & Bass, 2001).

Although our aim in developing MUSOS was to encourage replication of studies, an attempt to reproduce every paradigm used in music psychology would be too broad a design, and would thus reduce ease of use of the program. In selecting candidate paradigms for inclusion, we therefore first considered theories of long-term memory, and the ways in which memory has been studied and tested in music psychology as well as across domains, in order to design a framework that was sufficiently flexible to contain a selection of useful paradigms for non-musician researchers seeking to administer and replicate their own and others' studies.

Dual process models propose that recognition memory has two components, recollection, in which specific details of encountering an event or item may be retrieved, and familiarity, an awareness that one has encountered something before, but without the ability to retrieve further details (Jacoby, 1991; Yonelinas, 2002). Recognition may therefore be explicit, involving conscious recall of the event, or implicit, where an increased fluency or *priming* is demonstrated despite a lack of conscious awareness of retrieval (Schacter, 1987; Schacter & Church, 1992).

In memory studies, explicit retrieval is tested using two methods: recognition and recall (Schacter, 1987). Both methods involve presenting the participant with a list of items to study in an initial exposure phase. In recall studies, the participant is then asked to recall as many items as they can remember, in free or serial order. For a recognition study, the participant is presented with a combination of novel and earlier-presented items, and asked to identify those that they recognize from the exposure phase (Müllensiefen & Wiggins, 2011). Implicit memory studies differ from recognition studies in that the participant is not forewarned of the upcoming test during the exposure phase. Priming may be demonstrated experimentally in a variety of tasks such as word fragment and stem completion, lexical decision tasks, or picture completion (Schacter & Church, 1992).

The majority of paradigms testing both explicit and implicit memory (in general cognition studies as well as music) fall into a two or three-phase structure, in which the initial phase provides exposure to stimuli, the final phase tests memory for these stimuli, either through re-presentation of stimuli in implicit or explicit recognition studies, or providing a facility for the input of recalled items in recall studies. Manipulation of one or more factors under investigation may occur within the exposure phase, or during a second phase prior to testing. In music cognition, this has involved rating qualitative aspects of a piece of music such as similarity, familiarity or liking (Peretz, Gaudreau, & Bonnel, 1998), applying tempo or instrumentation changes (Halpern & Müllensiefen, 2008), or repeated exposure (Schellenberg, Peretz, & Vieillard, 2008).

In music, explicit recognition is one of the most commonly used methods for studying memory for musical items, due to the high level of experimental control possible (Sloboda & Parker, 1985). Studies of explicit recognition in music have yielded findings that musical key, timbre, tempo, and rhythmic content affect recognition of a melody (Halpern & Müllensiefen, 2008; Hébert & Peretz, 1997; Schellenberg & Habashi, 2015), that liking improves memory for music (Schellenberg et al., 2008; Stalinski & Schellenberg 2013), and that, as for other domains, distinctive content improves recognition (Bailes, 2010; Müllensiefen & Halpern, 2014; Schacter & Wiseman, 2006).

Although numerous studies of explicit recognition exist in the literature of music psychology (Müllensiefen & Wiggins, 2011), few studies of implicit memory for musical material have been conducted. One method developed by Warker and Halpern (2005) involved a musical adaptation of stem completion. In this study, following initial exposure, participants were presented with all but the final note of a group of previously heard and novel melodies, and were asked to complete the sequence by singing the most appropriate note to follow. This method differs from explicit recognition in that participants were not required to remember the note that followed, but were asked to judge which note would fit best musically (Warker & Halpern, 2005). Verification of the method as a test of implicit memory was demonstrated by Walker and Halpern using an encoding task to differentiate implicit memory for melodies, enhanced by shallow encoding of perceptual features, from explicit memory, which was found to be enhanced by deeper, semantic processing. Although promising, a search of the literature reveals that this method has not yet been replicated.

A further method used in the study of implicit memory for music involves exploiting the *mere-exposure effect* (Zajonc, 1968), in which liking for an item increases after exposure. This effect has been found to be particularly strong in music and may occur after a single reexposure (Peretz et al., 1998), persisting for up to 24 h (Stalinski & Schellenberg, 2013). The mere-exposure effect has therefore increasingly been used as an index of implicit memory for music (Halpern & O'Connor, 2000; Peretz et al., 1998). Implicit memory for music is shown by increased pleasantness ratings at test for items heard at exposure, in comparison to novel items (Halpern & Müllensiefen, 2008). Müllensiefen and Halpern (2014) further used this method to identify a dissociation in qualities of melodies that lead to improved implicit and explicit recognition.

Recall studies present a particular difficulty for those studying musical memory, as it has proven difficult to measure recall performance in music. Traditional methods have required the participant to use musical notation or to perform their response on a musical instrument (Deutsch, 1980) or by singing (Sloboda & Parker, 1985, Warker & Halpern, 2005). Müllensiefen and Wiggins (2011) discuss in detail the challenges presented when attempting to analyze data from sung responses, which they describe as "dirty" as a researcher must frequently make subjective judgments as to which note a participant intended to sing. A participant may be capable of perceiving pitch correctly, yet unable to exercise sufficient motor control over their vocal apparatus to sing their response in tune (Hutchins, Larrouy-Maestri, & Peretz, 2014). Responses that are a few cents above or below the note may be normalized with electronic equipment (see Warker & Halpern, 2005), but a singer with poor pitch control may miss the intended pitch by several semitones, or transpose segments of the melody while retaining correct pitch interval relationships (Dalla Bella, Giguère, & Peretz, 2007; Dalla Bella & Berkowska, 2009). Despite potentially possessing normal pitch perception, singers with such difficulties in vocal control are often excluded from studies, or the sample restricted to those with musical training (e.g., Levitin, 1994; Warker & Halpern, 2005). Although this may result in more reliable responses, this leaves researchers unable to investigate questions regarding untrained musicians, or to compare the effects of expert training in music with a control group. We provide with the MUSOS toolkit a computer-based method for participants to input recall responses, thus facilitating studies in untrained populations.

A further issue encountered by researchers wishing to study recall in music lies in the analysis of the data collected. Sung responses must be transcribed into musical or MIDI notation for analysis, requiring musical expertise on the part of the experimenter as well as participant (Müllensiefen & Wiggins, 2011). Unlike verbal recall, responses in the recall of musical melodies are rarely exact, and often involve partial recall of segments of the melodies, with errors or omissions of several notes. Scoring of musical recall data has therefore frequently involved subjective judgments as to how closely a response resembles the original (for an example, see Sloboda & Parker, 1985, p.157). Instead of using subjective musicological techniques in analysis, Müllensiefen and Wiggins (2011) recommend conducting the data analysis of such studies using computational tools capable of similarity

analysis, such as the SIMILE toolkit (Müllensiefen & Frieler, 2006), so that factors such as missing or distorted notes and transpositions may be taken into account. We therefore include with the MUSOS Toolkit a means of exporting recall data to CSV, along with a spreadsheet for analysis in Excel using the *edit distance*, or *Levenshtein distance* algorithm, a simple form of similarity analysis based on the number of edits needed to transform a participant's attempt into the original melody (Müllensiefen & Wiggins, 2011).

Rationale, aim, and scope of the present study

Although a considerable number of innovative studies continue to be contributed to the literature on music and memory, as for other domains, it is of concern that few replication studies are undertaken of both novel and existing experiments (Frieler et al., 2013). One possible reason for the lack of replication studies in music psychology may lie in the difficulty of measuring participant responses (Müllensiefen & Wiggins, 2011). Our aim was therefore to facilitate ease of administration and measurement, and thereby improve the replication of studies by music researchers, by providing an easy to use toolkit that is capable of reproducing a number of common paradigms.

The three-phase structure of exposure, manipulation and testing phases is common to a number of important studies across both music and general cognition. It is ideal for the construction of a toolkit that is easy for researchers to use. In terms of software design, the three phases may be used as a framework, within which modules for each phase may be selected and added to form test paradigms. For example, if testing the effects of repeated exposure on implicit memory for melodies, a module for exposure, re-exposure, and a final test of pleasantness ratings would be used. For explicit recognition, the re-exposure module would be removed, and the final module would be reconfigured to test recognition of old and new items. Although a number of noteworthy paradigms fall outside of this structure, it would not be possible to provide in a single program a means of replicating all past studies, nor would such a program be capable of being contained within a simple and thus usable architecture (John & Bass, 2001). Arguably many important studies that do not use a three-phase structure are already well replicated in the literature—for example, Deutsch's pitch comparison paradigm (Deutsch, 1970, 1972, 1973, 1974), Dowling's AB comparison method, used to present standard and comparison melodies for discrimination of changes in contour (Bartlett & Dowling, 1980; DeWitt & Crowder, 1986; Dowling, 1978; Dowling & Bartlett, 1981) and cohort theory studies using dynamic melody recognition (Bailes, 2010; Dalla Bella, Peretz, & Aronoff, 2003; Schulkind, 2004). In contrast, relatively few studies have been undertaken of musical recall and implicit memory for music (Müllensiefen & Wiggins, 2011).

We therefore aimed to use the three-phase structure to construct a modular framework that may be used for the study of recollection memory in music, covering implicit and explicit recognition and recall studies, to make it easier for researchers with or without musical training to administer and reliably measure studies in the general population, thus facilitating increased replication of both past and future studies.

We further aimed to provide with this software a novel, copyright-free set of stimuli that have been designed and tested according to musical properties known to be involved in recollection memory. In developing the stimuli accompanying the MUSOS Toolkit we first used Bailes's (2010) measures of the likelihood of occurrence of notes of the musical scale as a rule to compose melodies that were more, or less distinctive in content, and thus, more or less likely to be well remembered. We then verified these melodies by obtaining ratings from a group of pilot testers on the perceived distinctiveness and valence of the melodies, as variables associated with the likelihood of occurrence and memory for musical items (Bailes, 2010; Huron, 2006; Schmuckler, 1997).

We then used computer-based analysis to measure the properties of the full stimulus set, using FANTASTIC (Müllensiefen, 2009a) to compute data on pitch and intervallic predictability, tonality, and contour of the melodies. Within the stimulus set, we aimed to create two subsets of high- and low-difficulty melodies that researchers may use in testing. We selected those melodies that were highest and lowest in distinctiveness and valence, as rated by pilot testers, for use in a recognition study involving 26 participants. We further verified, using the data obtained from FANTASTIC that these subsets of melodies differ significantly in musical properties associated with the likelihood of remembering an item.

MUSOS Software

Software architecture and paradigm selection

Our aim in designing this toolkit was to provide a platform that would assist researchers to generate and reproduce studies of music and memory, regardless of their level of musical training. By using the two or three-phase structure common to memory paradigms across domains within a visual development environment (Max/MSP; Cycling '74, 2014a), we were able to construct a framework housing a system of modules that may be selected and inserted in a 'plug and play' fashion.

Our final selection of paradigms comprised explicit old/new recognition, implicit recognition (using the method described by Halpern and Müllensiefen (2008) as well as manipulation of the mere-exposure effect (Zajonc, 1968)), stem completion (following the method described by

Warker & Halpern, 2005), and free recall. To construct these paradigms, we provided five modules for exposure, rating of stimuli, recall, stem completion, and old–new recognition.

Software design

The main components of the MUSOS Toolkit are a Max/MSP *live.step* step-sequencer, used for the input and display of melodies, which is connected to a system of databases created from Max/MSP *jit.cellblock* objects. A step-sequencer is a device commonly used in popular electronic music production for the recording and automated playback of musical material. The sequencer *steps* through each division or beat of the musical bar, playing the note that is assigned to that beat (Aikin, 2014). In Max/MSP, the *live.step* object allows the user to interact with the sequencer via a grid interface, with notes represented as blocks within the grid. We chose this interface for use in MUSOS as it is intuitive to use, and does not require the participant or experimenter to be trained in reading a musical score. Each division of the *x*-axis of the grid represents a musical beat, with movement up and down the *y*-axis representing increases and decreases in pitch, respectively (see Fig. 1). Using this device, a melody may be represented as a series of coordinates (appearing as black blocks in Fig. 1) and stored in numerical form in a database for later retrieval and analysis.

The use of a step-sequencer enables participants (and experimenters) to easily compose melodies by adjusting the locations of blocks in the grid. In Max/MSP all musical cues, including note names, tempo, and beat divisions, may be removed from a step-sequencer (Cycling '74, 2014b), leaving a row of square blocks that the participant places into the desired position using the computer mouse. The participant does not need to be trained to identify notes on a chromatic keyboard, as the *y*-axis of the device is preset to the pitches of the modal scale used in the stimulus set using MIDI quantization (see Fig. 1). Advanced Max/MSP users may reconfigure the application to present custom scales using the documentation provided. This simple graphical interface is therefore easy for both trained musicians and untrained participants to use, and allows the variable of melody to be measured in isolation from rhythm, tempo, and timbre.

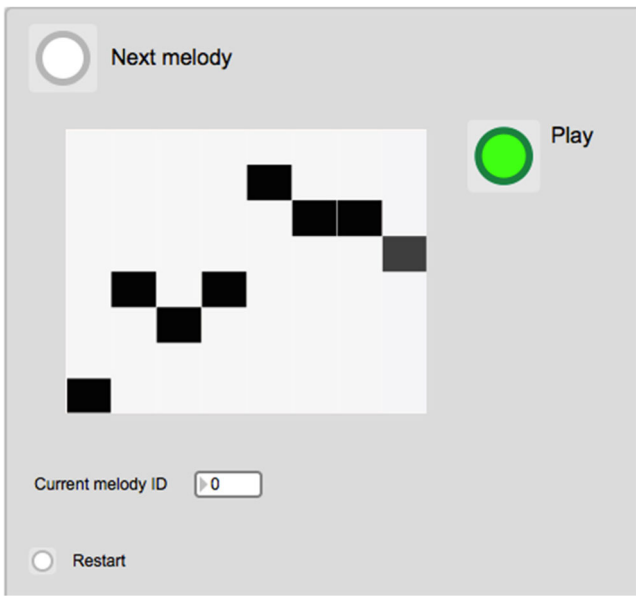

Fig. 1 The step-sequencer device used in the application. All visual cues, including note names and beat divisions, are removed, and the *y*-axis of the device is preset to a MIDI-quantized modal scale

Modules included in the software

Five modules are included with the MUSOS Toolkit. The *Exposure phase* module combines melodies from different conditions and displays these to the participant in random order. A *Rating* module allows participants to rate attributes of a selection of melodies (e.g., pleasantness, distinctiveness). Data from these ratings can be used for manipulation checks or correlational analyses (e.g., are melodies rated as more pleasant better remembered?), or to test the effect of repeated exposure to stimuli. Alternatively, this module may be added to the final test phase in order to measure implicit memory for items. The remaining three modules supplied with the application are also designed for experimental testing following exposure; these include the *Free Recall*, *Stem Completion*, and *Recognition* test modules.

Installation and connection of the modules to their databases is performed in Max/MSP Patching Mode. The experimenter then switches to Presentation Mode, in which the visual interface is displayed to the participant.

Free recall

The graphical interface of MUSOS is designed so that the responses of those with and without specialist training may be reliably recorded. In the Free Recall module provided with the MUSOS Toolkit, the participant is presented with a series of blank step-sequencers into which they may input as many melodies from the exposure phase as they are able to recall. The step-sequencer device allows an untrained participant to use a simple graphical interface to enter, listen to, and correct their response, thus ensuring that the data recorded are as close as possible to the participant's true response. Responses do not require normalization to the correct pitch, as the step-sequencer is preset to a MIDI-quantized scale. Since the Free Recall module records melodies to a Max/MSP jit.cellblock database, it may also be used as a standalone module to record and save new stimuli for use in the program.

Because interpretation of free recall data has also proven challenging for researchers (Müllensiefen & Wiggins, 2011; Sloboda & Parker, 1985), we provide with the MUSOS toolkit a method for computational analysis of free recall responses. In MUSOS, data are stored in numerical format, and may be exported to comma-separated value (CSV) format and converted for analysis with any suitable computational application. For researchers who are not familiar with such applications, we also provide an Excel spreadsheet for analysis of recall data in Excel using the *edit distance*, or *Levenshtein distance* algorithm, a simple form of similarity analysis based on the number of edits needed to transform a participant's attempt into the original melody (Müllensiefen & Wiggins, 2011). This method is capable of capturing subtle changes in response such as missing notes or transpositions of the melody without requiring subjective interpretation of the participant's intention.

An example of the output of free-recall analysis can be seen in Fig. 2. Participants' responses are listed in column A, with the original melodies in Column B. From Column D onward, each participant's entry is compared against the originals using the Levenshtein distance algorithm, which outputs values between 0 and 1, where 1 indicates a 100% match with the reference melody.

Unlike in verbal studies, recall responses in music are rarely exact, a common finding when working with both trained and untrained musicians (Müllensiefen & Wiggins, 2011). When using an algorithmic measure of musical similarity, a threshold is normally set above which matches between two melodies are considered unlikely to occur beyond chance, and are thus considered significant (Müllensiefen & Frieler, 2007). For edit distance analysis, Müllensiefen and Pendzich (2009) used a threshold of .46, although values of up to .6 are commonly used (Frieler, e-mail correspondence). On examination of the output of the edit distance analysis, values below .5 indicated poor correspondence with the original (see Fig. 2), so for the supplied examples a threshold of .6 was therefore set as an indication of memory beyond chance for the original melody.

F15 =Levenshtein3(A4,B15)/100

	A	B	C	D	E	F
1	RecallDB	ExposureDB	Melody ID	Attempt 1	Attempt 2	Attempt 3
2	65453342	47372737	LD6	0	0.12	0.12
3	04162164	33435434	HD3	0.38	0.12	0.75
4	33235433	57567023	LD14	0.25	0.12	0.12
5	20264611	60642231	HD1	0.25	0.12	0.12
6	45454322	20264611	LD7	0	0.12	0.12
7	13634543	00102310	HD14	0.12	0.38	0
8	01316340	24675310	LD4	0.12	0.12	0.12
9	60642021	43543232	HD4	0.38	0	0.38
10	01000110	13631454	LD3	0.12	0.25	0.38
11	35534221	43233210	HD2	0.12	0.12	0.38
12	33432100	01316340	LD5	0.25	0.38	0.12
13	24334333	21202012	HD8	0.12	0.12	0.12
14	47374727	65454342	LD11	0.88	0	0.25
15	75576567	24334433	HD10	0.25	0.12	0.5
16		37534230	LD15	0.25	0	0.38

D14 =Levenshtein3(A2,B14)/100

	A	B	C	D	E	F
1	RecallDB	ExposureDB	Melody ID	Attempt 1	Attempt 2	Attempt 3
2	65453342	47372737	LD6	0	0.12	0.12
3	04162164	33435434	HD3	0.38	0.12	0.75
4	33235433	57567023	LD14	0.25	0.12	0.12
5	20264611	60642231	HD1	0.25	0.12	0.12
6	45454322	20264611	LD7	0	0.12	0.12
7	13634543	00102310	HD14	0.12	0.38	0
8	01316340	24675310	LD4	0.12	0.12	0.12
9	60642021	43543232	HD4	0.38	0	0.38
10	01000110	13631454	LD3	0.12	0.25	0.38
11	35534221	43233210	HD2	0.12	0.12	0.38
12	33432100	01316340	LD5	0.25	0.38	0.12
13	24334333	21202012	HD8	0.12	0.12	0.12
14	47374727	65454342	LD11	0.88	0	0.25
15	75576567	24334433	HD10	0.25	0.12	0.5
16		37534230	LD15	0.25	0	0.38

Fig. 2 Sample output from a free-recall data analysis using the Levenshtein distance algorithm. Melodies are aggregated into eight-digit figures representing the eight degrees of the scale used in the melody. Each participant attempt in column A is compared against the original melodies in column B, to produce a matrix. Significant responses (>.6) are highlighted in red. In the top panel, two melodies with a Levenshtein distance of .5 contain a range of notes in common, but are otherwise not audibly similar. In contrast, the lower panel shows two melodies that have a Levenshtein distance of .88, which are almost identical with the exception of the fifth note

Further instructions for using the free recall analysis spreadsheets are provided in the MUSOS Toolkit documentation.

Stem completion

The Stem Completion module included with the MUSOS Toolkit is based on the method developed by Warker and Halpern (2005). Instead of requiring the participant to sing the most appropriate note to complete the melody, a computer-based interface is used. The module draws melodies from a task database that comprises a counterbalanced selection of items previously encountered in the Exposure phase alongside an equal number of novel melodies. The participant is presented with a step-sequencer containing all but the final note of a melody randomly selected from the task database. The participant first listens to the melody, and is then asked to select the note that would best follow by setting the final block in place. The result may be auditioned and corrected by the participant, if necessary, to ensure that the recorded melody reproduces their intended response (see Fig. 3). Although the present method involves completion of a single note, the module may be easily adjusted following the documentation provided by those fluent in the use of Max/MSP so that stem completion of two or four notes may be tested. Scoring of a stem completion study is considerably simpler than scoring the Free Recall task, as the melodies completed by the participant must simply be exported to CSV format and compared to the original versions, which are stored by the program in a separate database. A matching final note is scored as a correct response, and all other responses are scored zero (Warker & Halpern, 2005). An Excel spreadsheet is also provided with the MUSOS Toolkit for scoring of Stem Completion data, along with sample data from an eight-note stem completion study.

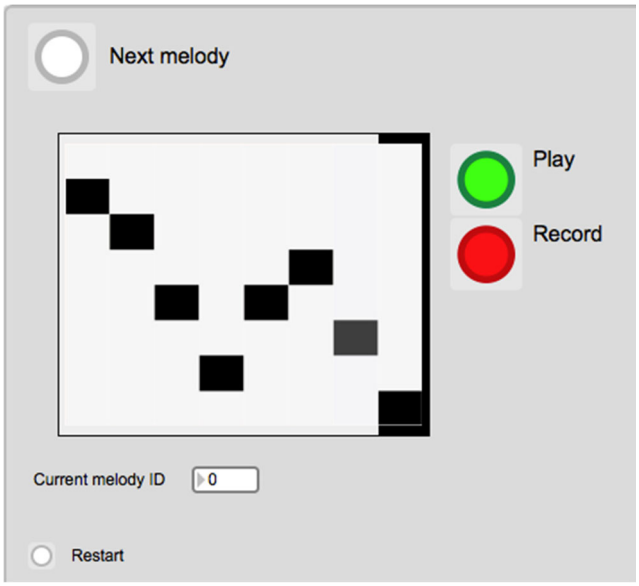

Fig. 3 The Stem Completion module, based on the method developed by Warker and Halpern (2005). The participant is presented with all but the last note of the melody and is asked to complete the melody with the most appropriate final note, by adjusting the block in the section outlined in black

Recognition

The Recognition module provided with MUSOS uses a simple listwise recognition procedure, similar to those used in verbal and facial recognition studies (Müllensiefen & Wiggins, 2011). The module retrieves melodies in random order from the recognition task database, again comprising an equal number of melodies previously encountered in the exposure phase, counterbalanced with novel melodies. (When configuring the application to test both stem completion and recognition, the exposure phase melodies may be assigned in counterbalanced order to the two modules, so that no duplicates occur.) The Recognition module differs from the others as the step-sequencer interface is removed and replaced with a progress bar, in order to ensure that participants do not rely on the visual features of the sequencer for recall. Participants listen to each melody in turn, and use a dial-based control to input their response to the statement, "I heard this melody in the previous task." Responses are recorded on a scale from +3 to –3, where +3 indicates *strongly agree*, 0 indicates *neither agree nor disagree*, and –3 indicates *strongly disagree* (see Fig. 4).

Rating and the mere-exposure effect

The Rating module simply retrieves melodies from a task database and presents them to the participant alongside a dial based input for ratings using the same scale as used in the Recognition module. The basic module presents melodies using the step-

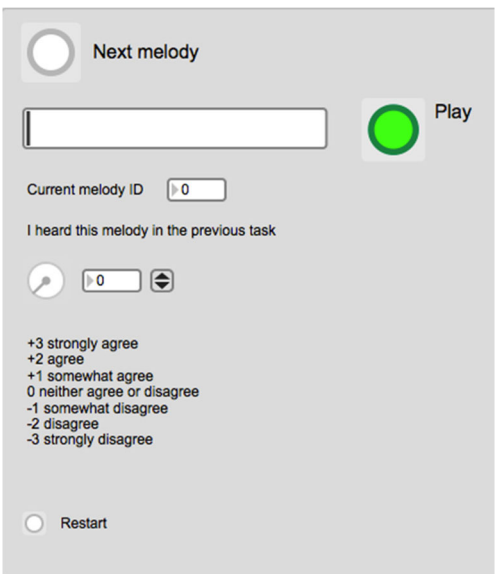

Fig. 4 The Recognition module. The step-sequencer interface is removed and replaced by a progress bar. A dial is provided for participants to input the degree to which they recognize the item

sequencer. An alternate form of the Rating module (*RecognitionImplicit*) is used for testing implicit memory via the *mere-exposure effect* (Zajonc, 1968), in which liking for an item increases after exposure (Peretz et al., 1998; Stalinski & Schellenberg, 2013). The RecognitionImplicit module uses the same progress bar as the Recognition module, to avoid visual recognition of melodies from the step-sequencer. Because this method also requires a measurement of liking for melodies at initial exposure in order to detect increases in liking corresponding with repeated exposure, a modified form of the Exposure module, *Exposure-Rating*, is provided that incorporates the same rating mechanism on-screen.

Alternate configurations of the dial component of the Rating module are available to advanced Max/MSP users. Configuring the dial to a range of three steps instead of seven would make analysis of remember/know judgments (Tulving, 1985) possible, by instructing the participant to record a remember judgment with a value of 0, know with a value of 1, and guess with a value of 2.

Stimulus development

The total stimulus set comprises 156 melodies, 78 of eight-note length and 78 of 16-note length. Below we describe the process of construction of the melodies. We then present the results of two pilot tests conducted to establish properties of the stimuli. The first provided data on the properties of each melody, including subjective ratings of distinctiveness and valence, and computational analysis of pitch, intervallic, tonal, and contour information using the software FANTASTIC (Müllensiefen, 2009a). The second test identified a subset of melodies that varied in musical properties affecting difficulty of recognition (i.e., one set of relatively difficult to recognize melodies and one set of easy to recognize melodies), which were then tested in a recognition study involving 26 participants.

Stimulus properties

Scale

In providing an original stimulus set, we aimed to ensure that the tonality of the melody was unfamiliar to Western listeners, thus minimizing the chances that a novel melody presented during an experiment will remind the listener of some other melody previously heard outside of the laboratory. The melodies were therefore composed using a seven-note scale commonly used in world music (*Maqam Kurd*, in Arabic music, also known as the *Phrygian mode* in Western medieval music, and as *Hanumatodi rāgam* in Carnatic music). This scale is structured around a semitone–tone–tone–tone pentachord followed by a semitone–tone–tone tetrachord (concluding on the upper octave), which differs in structure from both the Western major and minor scales (see Fig. 5 for a comparison against these scales).

Note name	D	E ♭	F	G	A	B ♭	C	D
MIDI note	62	63	65	67	69	70	72	74
Major scale	D	E*	F#*	G	A	B*	C#*	D
Minor	D	E*	F	G	A	B ♭	C#*	D

Fig. 5 The scale used in the MUSOS Toolkit is provided in the top row in musical note names, and on the second row as MIDI note numbers. The scale is then compared to the major and minor scales of Western music on the third and fourth rows. Asterisks indicate notes that differ from those in the major and minor scales

All stimuli are composed in 4/4 meter and are isochronic in rhythm, with four quarter notes per bar. Although rhythm is also important in the study of musical memory, in developing these stimuli we chose to focus on those aspects of melody (pitch, interval, tonality, and contour) that may cause a melody to be easy or difficult to remember (Deutsch, 1975, 1980; Dowling, 1978; Krumhansl, 1991; Schmuckler, 1997). Isochronic melodies are commonly used in such studies in which the focus is on aspects of melody that affect memory for music (e.g., Halpern & Bower, 1982). Advanced Max/MSP programmers may adjust the *live.step* sequencer to present their own melodies using varied rhythm.

Tonality

To ensure sufficient variety within the melody collection, the stimuli were permitted to begin or end on any of the eight notes of the scale. A possible $8^8 = 16.7$ million sequences can be generated from an eight-note melody composed on an eight-note scale, and $2.81e^{+14}$ for a 16-note melody on the same scale; thus, sufficient degrees of freedom were available within this structure to eliminate the possibility that the stimuli were too similar.

Because Western modal scales consist of identical intervallic structures, varying only by the note on which they begin (the Ionian mode being identical to the modern major scale), permitting the melodies to begin on any note of the scale meant that the melodies varied in the degrees to which they conformed to Western concepts of tonality. Further analysis was conducted in the tests below by using FANTASTIC to assess the implicit tonality of the melody, and the *tonalness*, or degree to which the melody correlated to a given scale (Müllensiefen, 2009b).

Stimulus distinctiveness

In a study of the role of distinctiveness in online recognition of melodies, Bailes (2010) used the Humdrum toolkit (Huron,

1994) to calculate the distinctiveness of scale degree and intervallic information, finding that stepwise intervals of a major second have a higher probability of occurring in Western melodies, and are thus more typical than less frequently occurring wider intervals such as the augmented fourth. In the same study, bit values were also computed indicating the relative probability of a scale degree occurring in a melody. This information was used as a guide for composition of the MUSOS stimulus set, with melodies designed to be highly distinctive including wider intervals and less frequently occurring scale degrees, whereas melodies designed to be more typical (i.e., low distinctiveness) were composed with regularly occurring notes of the scale, and stepwise passages.

Stimulus valence

Although a non-Western scale was used for the experiment, the majority of participants were of Western origin, and would therefore have acquired Western constructs of consonance and dissonance through passive listening experiences (Johnson-Laird, Kang, & Leong, 2012; Levitin & Tirovolas, 2010). Therefore, when composing melodies expected to be perceived as high or low in valence, Western musicological constructs of consonance and dissonance were used, with dissonant intervals based on augmented and diminished intervals and chords included in the low-valence melodies, and consonant intervals based on major or minor chords and their inversions in the high-valence melodies (Johnson-Laird et al., 2012).

Pilot test 1: Distinctiveness and valence ratings

Because composition according to computer-calculated values and musicological principles may not always reflect the perception of individual listeners, the set of stimuli were rated by a group of pilot testers for values of distinctiveness and valence.

Method

Thirty-six participants were recruited to take part in a Web-based experiment. Those who were first-year students of the University of Tasmania School of Psychology received course credit for participation; the remainder were entered into a draw to receive vouchers as remuneration.

Melodies were presented to participants in one of four randomized orders, with eight-note melodies presented in the first block of testing, and 16-note melodies presented in the second block. Within each block, the group of melodies was divided into four sections. Participants were instructed to take a brief break before proceeding to the next page. Participants were asked to listen carefully to each melody and rate two accompanying statements, "This melody has distinctive features," and "This melody is likeable." Responses were recorded on a Likert-type scale ranging from –3 to +3, where –3 indicated *strongly disagree*, 0 indicated a neutral response, and +3 indicated *strongly agree*.

Results and discussion

Raw values of distinctiveness and valence for each melody were summed across all participants. For eight-note melodies, the mean distinctiveness ratings ranged from –0.14 to 1.25 ($M = 0.48$, $SD = 0.30$). Total scores for each melody were then converted to z scores, which ranged from –2.06 to 2.56. The mean valence ratings for eight-note melodies ranged from –0.58 to 1.25 ($M = 0.17$, $SD = 0.34$), which when converted to z scores revealed a range of –2.18 to 3.12. For 16-note melodies, the mean distinctiveness ratings ranged from –0.17 to 1.31 ($M = 0.47$, $SD = 0.29$), with z scores ranging from –2.22 to 2.93. The mean valence ratings ranged from –0.50 to 1.19 ($M = 0.23$, $SD = 0.31$), with z scores ranging from –2.36 to 3.12.

The full set of scores for each melody is provided with the MUSOS stimulus set (available for download at http://www.soundinmind.net/MUSOS/MUSOS.zip).

Computational analysis of the stimulus set

We computed feature summary statistics and m-type summary statistics of the melodies using FANTASTIC (Müllensiefen, 2009a). Features included pitch range, variance (standard deviation) and entropy, intervallic range, mean interval, and intervallic variance (standard deviation) and entropy. Information on tonality, including the mode of each melody (major or minor scale) and the degree to which the melody correlated with the identified scale, was also computed. Finally, the calculations included several methods of describing the contour of each melody, including Huron's (1996) eight contour types, as well as interpolation, polynomial, and step contour. Further descriptions of these statistics and the calculations by which they may be obtained are available in the FANTASTIC documentation (Müllensiefen, 2009b). The full set of statistics describing each melody is included in a spreadsheet accompanying the stimulus set.

We then conducted Bayesian correlations between the computed features of the melodies and participant ratings of distinctiveness and valence, in order to examine whether the computational analysis showed a relationship to participant ratings. Table 1 presents Bayes factors and Pearson correlation values for participant ratings of distinctiveness and valence. According to Jeffreys's (1961) criteria, Bayes factors of 3 or above represent substantial evidence, and Bayes factors of 10 or above represent strong evidence for the hypothesis that the variables were correlated.

Table 1 Bayesian correlations between features of melodies and participant ratings of the melodies' distinctiveness and valence

Variable		Distinctiveness (Mean Rating)	Valence (Mean Rating)
Valence	*r*	.57	–
	BF_{10}	$4.301e^{+15**}$	
Pitch range	*r*	.41	.01
	BF_{10}	162,012.10**	0.10
Pitch entropy	*r*	.25	.14
	BF_{10}	13.16**	0.45
Pitch standard deviation	*r*	.43	–.02
	BF_{10}	538,776.46**	0.10
Interval absolute range	*r*	.25	–.22
	BF_{10}	11.33**	4.63*
Interval absolute mean	*r*	.42	–.03
	BF_{10}	330,634.50**	0.11
Interval absolute standard deviation	*r*	.25	–.25
	BF_{10}	11.03**	12.83**
Interval mode	*r*	.40	.24
	BF_{10}	48,000.32**	7.42*
Intervallic entropy	*r*	.37	.12
	BF_{10}	9,500.71**	0.32
Tonalness	*r*	.25	.23
	BF_{10}	14.36**	5.96*
Tonal clarity	*r*	–.22	–.06
	BF_{10}	4.68*	0.13
Tonal spike	*r*	.02	–.04
	BF_{10}	0.10	0.12
Interpolation contour mean gradient	*r*	.07	–.17
	BF_{10}	0.15	1.03
Interpolation contour standard deviation	*r*	.09	–.16
	BF_{10}	0.18	0.77
Interpolation contour direction change	*r*	–.06	–.07
	BF_{10}	0.13	0.14
Step contour global variation	*r*	.43	–.02
	BF_{10}	552,740.64**	0.10
Step contour global direction	*r*	–.02	.04
	BF_{10}	0.10	0.11
Step contour local variation	*r*	.42	–.02
	BF_{10}	301,586.70**	0.10
Polynomial coefficient 1	*r*	–.13	–.13
	BF_{10}	0.34	0.34
Polynomial coefficient 2	*r*	–.08	–.08
	BF_{10}	0.15	0.15
Polynomial coefficient 3	*r*	.15	.15
	BF_{10}	0.61	0.61

* indicates substantial support for the hypothesis; ** indicates strong support for the hypothesis.

Significant positive correlations were found between participant ratings of distinctiveness and variables describing pitch, intervallic, and tonal content, with weak to moderate effects. Thus, as range and variability in pitch and intervallic content increased, melodies were more likely to be perceived as distinctive rather than typical.

This relationship is consistent with Bailes's (2010) calculations of distinctive pitch and intervallic content, which were used in composition of the melodies.

A weak-to-moderate correlation between distinctiveness and tonalness, or the degree to which a melody correlated with the Western major or minor scales, was observed. However, Temperley's (2007) statistic of *tonal clarity* showed a weak negative correlation with distinctiveness. This statistic describes the ratio between the first and second highest correlations with a Western major or minor key. Higher values indicate closer correlations with a single, rather than several, keys (Temperley, 2007); therefore, a negative correlation with tonal clarity indicates that melodies that were more ambiguous in tonality were perceived as more distinctive. Since the tonal clarity statistic is based on the probability of a key given the pitch class set of the melody (Temperley, 2007), this finding again shows consistency with Bailes's (2010) calculations of distinctive and typical notes of the major and minor scale, used in composition of the melodies. This result is further consistent with Vuvan and colleagues' (2014) findings of a distinctiveness effect in memory for highly unexpected musical tones.

Regarding the contour of melodies, only global and local variation in step contour were related to distinctiveness. Step contour describes a curve drawn by plotting duration against pitch; thus, the moderate positive correlations found here indicate that melodies containing greater variety in contour were rated as more distinctive.

Although participant ratings of distinctiveness and valence showed a moderate positive correlation, fewer of the computed statistics describing the melodies were related to valence. Intervallic range (the difference between the maximum and minimum interval) and standard deviation were negatively related to valence; thus, melodies with less variation in intervallic content were perceived as higher in valence. However, a wider modal (most frequent) interval also predicted higher valence. Tonalness was also positively correlated with valence. This result is consistent with Huron (2006) and with Johnson-Laird and colleagues' (2012) study of the perception of pleasantness in consonant and dissonant chords. As for distinctiveness, a relationship may again be observed between correlations with valence and the rules on which composition was based, where dissonant augmented and diminished intervals were used to compose melodies low in valence, whereas consonant intervals of fourths, fifths, and major and minor thirds and sixths were used frequently to compose melodies planned to be high in valence.

Pilot test 2: Difficult versus easy to recognize stimuli

A brief recognition test was conducted to establish a subset of melodies from the stimulus set for use as test items designed to be either very easy or very difficult to remember. According to Rajaram's (1996) distinctiveness–fluency framework, distinctive items are more readily identified in a test of explicit recognition, a finding that has been replicated across visual, verbal, and musical domains (Bailes, 2010; Brandt, Gardiner, & Macrae, 2006; Bülthoff & Newell, 2015; Cohen & Carr, 1975). Thus, as a starting point for identifying a set of easy and difficult to recognize items, we chose a group of melodies from the stimulus set with very high values of distinctiveness (which should be relatively easy to recognize) and a set with very low values of distinctiveness (which should be difficult to recognize). We further used the values obtained through analysis using FANTASTIC to identify musical properties on which the easy- and difficult-to-recognize melodies differed significantly.

Method

Participants were 26 first-year Psychology students (three males, 23 females) at the University of Tasmania who received course credit for participation. Participants were not required to have received training in music.

The MUSOS application was configured to present participants with two recognition tests, one using the eight-note melodies, and the other using 16-note melodies. Two pairs of Exposure and Recognition modules were used for this design. Test administration was counterbalanced by creating two versions of the application, the first presenting participants with the eight-note test first, and the second with the 16-note test first.

Forty-eight melodies from each of the eight- and 16-note melody collections were selected as stimuli for inclusion in the pilot test. In each note-length category, the 24 melodies with the highest and lowest ratings of distinctiveness constituted the low-difficulty and high-difficulty stimuli, respectively.

Procedure

Participants were randomly assigned to complete either the eight-note recognition test or the 16-note recognition test first. Participants were given brief instructions on how to use the software by the experimenter, and then proceeded to operate the program in a self-directed manner. In the exposure phase of each experiment, participants were presented with the 24 melodies in random order, and were asked to listen carefully to each of the melodies. Then, for the recognition test, participants were presented with the 24 previously heard and 24 novel melodies in random order. Participants were asked to rate whether they thought that the melody was one they had previously heard in the exposure phase, or a novel melody, according to the statement "I heard this melody in the previous task," where +3 indicated *strongly agree* and –3 indicated *strongly disagree*.

Results and discussion

Using the spreadsheets provided with the MUSOS Toolkit, randomization was removed and participant ratings were calculated for low- and high-difficulty melodies when presented as targets during the exposure phase and when appearing as lures (i.e., when the melody did not appear in the exposure phase). From these values, total ratings for targets and lures for low- and high-difficulty melodies of each note length were calculated.

Following the initial analysis, we discovered that some of the melodies selected were not performing as would be expected according to the values obtained in the first pilot test. We examined participants' mean recognition ratings for each melody. In both the eight- and 16-note melody collections, we removed four melodies from both the low- and high-difficulty categories that were most likely to be rated as being earlier presented when in fact they had not been. We then ran the following analyses on the final set of 80 melodies (20 low-difficulty and 20 high-difficulty melodies in each note-length category), with the aim of establishing a reliable stimulus set of high- and low-difficulty melodies that could be used by researchers for testing with the MUSOS Toolkit.

The mean ratings for eight- and 16-note melodies appearing as targets and lures are given in Table 2. The data for the final collection of melodies were analyzed with a 2 (Condition: Target, Lure) × 2 (Difficulty: low, high) × 2 (Length: eight-note, 16-note) repeated measures analysis of variance (ANOVA). A large and statistically significant main effect of condition, $F(1, 25) = 25.34$, $p < .001$, $\eta_p^2 = .50$, indicated that participants could distinguish target melodies from lures, evidenced by higher ratings for targets than lures. This indicated that participants could distinguish target melodies from lures overall (i.e., collapsing across different level of difficulty and length).

Table 2 Participant ratings for melodies appearing as targets and lures

Condition	Difficulty		
	Low	High	Overall
8-Note Melodies			
Targets	6.3 (6.1)	1.3 (6.3)	3.8 (4.7)
Lures	–0.6 (7.8)	–1.1 (6.2)	–0.8 (5.9)
16-Note Melodies			
Targets	10.8 (6.6)	1.2 (8.3)	6.0 (6.1)
Lures	0.2 (6.5)	0.9 (8.2)	0.5 (5.4)
Overall			
Targets	8.6 (5.4)	1.2 (5.6)	4.9 (4.3)
Lures	–0.2 (5.2)	–0.1 (6.2)	0.2 (4.9)

Figures shown in parentheses indicate standard deviations.

For establishing the effect of difficulty, the critical result was a large and significant two-way interaction between difficulty and condition, $F(1, 25) = 16.05$, $p < .001$, $\eta_p^2 = .39$, indicating that participants' ability to distinguish target melodies from lures varied depending on difficulty. Simple-effects analyses (using a Bonferroni-corrected alpha level of .006) showed that participants were much better at distinguishing targets from lures with the low-difficulty than with the high-difficulty melodies. For low-difficulty melodies, higher ratings were given to targets than to lures for both eight-note melodies, $t(25) = 3.82$, $p = .001$, 95% CI [3.19, 10.67], $d = 0.99$, and 16-note melodies, $t(25) = 6.86$, $p < .001$, 95% CI [7.48, 13.90], $d = 1.63$. In contrast, for high-difficulty melodies we found little difference in the ratings given to targets and lures for both eight-note melodies, $t(25) = 1.50$, $p = .146$, 95% CI [–0.88, 5.65], $d = 0.38$, and 16-note melodies, $t(25) = 0.12$, $p = .905$, 95% CI [–4.33, 4.870, $d = 0.03$.

Further exploratory analysis revealed that the advantage for low-difficulty melodies emerged because low-difficulty target melodies were easier to recognize, rather than because low-difficulty lures were easier to reject. For targets, higher ratings were given to low-difficulty than high-difficulty melodies of eight-note, $t(25) = 3.21$, $p = .004$, 95% CI [1.82, 8.33], $d = 0.82$, and of 16-note, $t(25) = 5.63$, $p < .001$, 95% CI [6.14, 13.24], $d = 1.29$, length. For lures, there was little difference in ratings between low and high difficulty for eight- or 16-note melodies (all t values < 1).

Together, the results indicate that recognition performance was better for the low-difficulty melodies than for the high-difficulty melodies, and that this applied for eight-note and 16-note melodies.

Computational analysis of low- and high-difficulty melodies

We conducted independent-samples Bayes factor t tests, using the default prior (.707) to identify those variables on which the high- and low-difficulty melodies differed significantly. We included in this analysis both the participant ratings of distinctiveness and valence, and all variables measured using FANTASTIC. Since the recognition testing had demonstrated that performance was better for low-difficulty melodies in both the eight- and 16-note melodies, we collapsed the data to include the eight- and 16-note melodies together in the low- and high-difficulty data sets.

Table 3 presents descriptive statistics and Bayes factors for the melodies. According to Jeffreys's (1961) criteria, Bayes factors above 3 represent substantial support for the hypothesis, and Bayes factors of 10 or above represent strong evidence. As is evident, there were significant differences for these particular variables (i.e., moderate or higher support was obtained for the hypothesis that the two groups of melodies differed).

Table 3 Bayes factor *t* tests and descriptive statistics for low- and high-difficulty melodies

Variable	BF_{10}	Error%	Difficulty	Mean (*SD*)
Distinctiveness (mean rating)	$8.25e^{+29**}$	$4.12e^{-37}$	High	0.14 (0.12)
			Low	0.81 (0.17)
Valence (mean rating)	$3.97e^{+6*}$	$3.29e^{-12}$	High	0.01 (0.20)
			Low	0.46 (0.37)
Pitch range	366.48^{**}	$8.99e^{-8}$	High	6.76 (2.99)
			Low	9.40 (2.53)
Pitch entropy	4.68^{*}	$1.05e^{-5}$	High	0.42 (0.09)
			Low	0.47 (0.07)
Pitch standard deviation	313.95^{**}	$1.12e^{-7}$	High	2.46 (1.11)
			Low	3.45 (0.99)
Interval absolute range	1.27	$1.15e^{-5}$	High	4.94 (3.21)
			Low	6.31 (2.98)
Interval absolute mean	62.83^{**}	$9.78e^{-7}$	High	2.53 (1.26)
			Low	3.70 (1.55)
Interval absolute standard deviation	0.95	$3.37e^{-5}$	High	1.88 (1.30)
			Low	2.39 (1.21)
Interval mode	$1{,}111.91^{**}$	$2.21e^{-8}$	High	3.06 (1.33)
			Low	4.61 (1.67)
Intervallic entropy	440.21^{**}	$6.96e^{-8}$	High	0.46 (0.07)
			Low	0.53 (0.07)
Tonalness	7.99^{*}	$7.93e^{-6}$	High	0.63 (0.10)
			Low	0.69 (0.11)
Tonal clarity	5.16^{*}	$1.01e^{-5}$	High	1.20 (0.13)
			Low	1.13 (0.10)
Tonal spike	0.23	$2.15e^{-4}$	High	0.19 (0.03)
			Low	0.19 (0.02)
Interpolation contour mean gradient	0.28	$2.20e^{-4}$	High	2.64 (2.35)
			Low	2.97 (2.17)
Interpolation contour standard deviation	0.29	$2.19e^{-4}$	High	2.90 (3.00)
			Low	3.39 (3.16)
Interpolation contour direction change	0.25	$2.19e^{-4}$	High	0.42 (0.37)
			Low	0.38 (0.36)
Step contour global variation	315.70^{**}	$1.11e^{-7}$	High	2.32 (1.05)
			Low	3.25 (0.93)
Step contour global direction	0.24	$2.16e^{-4}$	High	–0.05 (0.40)
			Low	–0.07 (0.40)
Step contour local variation	54.40^{**}	$1.18e^{-6}$	High	0.29 (0.14)
			Low	0.41 (0.17)
Polynomial coefficient 1	0.72	$7.10e^{-5}$	High	0.28 (1.75)
			Low	–0.62 (3.03)
Polynomial coefficient 2	1.38	$7.70e^{-6}$	High	0.33 (1.96)
			Low	–0.76 (2.76)
Polynomial coefficient 3	1.69	$3.65e^{-6}$	High	–0.22 (0.68)
			Low	0.22 (1.12)

* indicates substantial support for the hypothesis; ** indicates strong support for the hypothesis.

Low-difficulty melodies were higher in perceived distinctiveness and valence, as well as in pitch range, pitch standard deviation, and pitch entropy. Low-difficulty melodies also had a higher interval absolute mean and a wider interval mode, and

were higher in interval entropy. Overall, these melodies could therefore be said to contain greater variation in intervallic content. An advantage for melodies with more distinctive pitch and intervallic content is consistent with Bailes's (2010) findings regarding the role of distinctive material in the point of recognition of a melody. Tonalness in low-difficulty melodies was higher, which shows consistency with Deutsch's (1970, 1972, 1973) and Krumhansl's (1979, 1991) studies demonstrating the role of scale and tonal relationships in facilitating memory for melodies. However, low-difficulty melodies were also lower in tonal clarity—that is, were more ambiguous in key—and may thus have facilitated recognition as less expectable events (Schmuckler, 1997; Vuvan et al., 2014). Interpolation contour did not differ between the two groups, but step contour global and local variation was higher in the low-difficulty melodies; thus, greater variation in contour was associated with improved recognition. This finding is consistent with Dowling (1978); (Bartlett & Dowling 1980; Dowling & Bartlett, 1981) and Halpern's (1984) studies demonstrating that similar contours are highly confusable, whereas variation in contours improves discrimination.

In summary, melodies that were easier to recognize can be described as containing greater variety in pitch and intervallic content, wider intervals and greater pitch range, and greater variation in contour. In contrast, difficult-to-recognize melodies had less variation in pitch and intervallic content and were more uniform in contour. Low-difficulty melodies also correlated more closely with Western musical scales and were more likely to correlate with a single rather than multiple tonalities. These variables associated with improved recognition of melodies were also shown, in the analysis of the full stimulus set above, to be associated with an increase in the perceived distinctiveness and valence of the items, further verifying the procedure involved in composing a set of high- and low-difficulty melodies.

The results of recognition testing, together with computational measurement of the melodies, verified the classification of a group of stimuli from the accompanying stimulus set into a prepackaged set of hard- and easy-to-recognize items that researchers may then use for testing any of the paradigms supplied with the MUSOS Toolkit.

Conclusion

In this article, a computer-based application was presented that is designed to facilitate the ease of administration and replication of studies of explicit and implicit memory for music. The application was designed with the aim of addressing two practical methodological issues that may be hindering replication studies in music psychology, an emerging field in which important findings have been made but replication rates are low (Frieler et al., 2013), specifically, difficulties in measuring recall responses and the availability of novel stimuli. The results of pilot testing with a sample of undergraduate students demonstrated that the software can be used easily by participants, and established some important characteristics of melodies in the accompanying stimulus set.

One advantage presented by a computer-based method is that it may be used for testing in the general population, whereas traditional methods involving instrumental performance or singing require trained musicians as test administrators as well as participants. The MUSOS application is easy for non-musicians to use, as demonstrated during testing in which all participants were able to use the program in a self-directed manner with minimal instruction. The modular basis of the software means that a researcher with or without musical training may develop and administer tests investigating memory for musical items.

This method, although practical and easy for researchers without expert musical training to use, is by no means a panacea for understanding the full complexity of musical recall responses. The limited number of studies conducted to date into the free recall of music clearly indicates that further research is needed before we have a complete understanding of musical memory (Müllensiefen & Wiggins, 2011). The MUSOS Toolkit is intended to provide researchers with the means to build an evidence base supporting our understanding of music cognition, so that we may investigate with greater reliability the free recall of melodies, and, using the accompanying stimulus set of hard and easy to recognize melodies, replicate stem completion studies such as that of Warker and Halpern (2005), or studies of implicit memory for music such as those by Halpern and Müllensiefen (2008). The source code of the MUSOS application, its accompanying documentation and stimulus set are made freely available to researchers who may wish not only to contribute such evidence through the replication of existing studies, but also to create conceptual replications, in which properties of the original study are varied or extended. Although exact replications are important initially to verify that a theory may be supported, conceptual replications test the extent to which a theory may be generalized across differing conditions (Frieler et al. 2013).

One further limitation that must be acknowledged is that a single toolkit cannot be capable of replicating every historic study of music and memory. Developing a full understanding of the factors involved in memory for music is a complex undertaking. Certainly, some factors cannot be understood without the need for novel and unique paradigms, which could not easily be included within a modular framework. However, as mentioned earlier, many of these important paradigms are already well replicated in the literature. Deutsch developed her pitch comparison paradigm for a series of studies investigating the pitch memory store (Deutsch, 1970, 1972, 1973, 1974; reviewed in Deutsch, 1975), which were extended by Krumhansl (1979) to build a model describing the role of

harmonic relationships in pitch memory. More recently, Mavromatis and Farbood (2012) used the same procedure to investigate the harmonic context of the comparison tone. It is noteworthy that all of these studies have involved electronic administration rather than human performance. Dowling's studies of the differential storage of scale and contour (Bartlett & Dowling, 1980; Dowling, 1978; Dowling & Bartlett, 1981) were replicated using electronic software in a series of studies by DeWitt and Crowder (1986). Extensive study has been undertaken of cohort theory in the storage and retrieval of melodies using dynamic melody recognition paradigms (Bailes, 2010; Dalla Bella et al., 2003; Schulkind, 2004), whereas there remains a pressing need to facilitate reliable studies of free recall (Müllensiefen & Wiggins, 2011).

Because MUSOS is easy to use and to configure, the requirement for expert musical training on the part of the researcher can be avoided. By providing participants with an accessible computer-based interface, this application resolves issues with "dirty" raw data captured through sung responses (Müllensiefen & Wiggins, 2011), and contributes further to the standardization of testing in this field, which Müllensiefen and Wiggins proposed may be addressed through the use of computer technology. The importance of extending research participation to the general population, rather than those who are reliably able to sing in tune, cannot be understated; if untrained musicians continue to be excluded from studies, the results cannot be said to generalize to an understanding of music perception, because it has already been demonstrated that trained musicians listen to music differently from those without training (Mikutta, Maissen, Altorfer, Strik, & Koenig, 2014).

Author note This research was funded by Australian Research Council Discovery Grant DP140103746 to M.A.P. and others, and an Australian Government Research Training Program Scholarship to M.R.

References

Aikin, J. (2014). Step sequencing basics. *Electronic Musician, 30,* 80–82.

Aruffo, C., Goldstone, R. L., & Earn, D. J. D. (2014). Absolute judgement of musical interval width. *Music Perception, 32,* 184–198. doi:10.1525/MP.2014.32.2.184

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

Baddeley, A. D., & Hitch, G. J. (1974). Working memory. In G. H. Bower (Ed.), *The psychology of learning and motivation: Advances in research and theory* (Vol. 8, pp. 47–89). New York, NY: Academic Press. doi:10.1016/S0079-7421(08)60452-1

Bailes, F. (2010). Dynamic melody recognition: Distinctiveness and the role of musical expertise. *Memory & Cognition, 38,* 641–650. doi:10.3758/MC.38.5.641

Bartlett, J. C., & Dowling, W. J. (1980). Recognition of transposed melodies: A key-distance effect in developmental perspective. *Journal of Experimental Psychology: Human Perception and Performance, 6,* 501–515. doi:10.1037/0096-1523.6.3.501

Berz, W. L. (1995). Working memory in music: A theoretical model. *Music Perception, 12,* 353–364. doi:10.2307/40286188

Brandt, K. R., Gardiner, J. M., & Macrae, C. N. (2006). The distinctiveness effect in forenames: The role of subjective experiences and recognition memory. *British Journal of Psychology, 97,* 269–280. doi:10.1348/000712605X73685

Bülthoff, I., & Newell, F. N. (2015). Distinctive voices enhance the visual recognition of unfamiliar faces. *Cognition, 137,* 9–21. doi:10.1016/j.cognition.2014.12.006

Cohen, M. E., & Carr, W. J. (1975). Facial recognition and the von Restorff effect. *Bulletin of the Psychonomic Society, 6,* 383–384. doi:10.3758/BF03333209

Creative Commons (2013). *Attribution-Non-Commercial-ShareAlike 4.0 international license*. Retrieved 4 February, 2015, from http://creativecommons.org/licenses/by-nc-sa/4.0/legalcode

Cycling '74 (2014a). Max/MSP 6.1 [Computer software]. Walnut, CA: Author. Retrieved from https://cycling74.com

Cycling '74 (2014b). Max 6 help and documentation. Retrieved 7 February, 2015, from https://cycling74.com/docs/max6/dynamic/c74_docs.html

Dalla Bella, S., & Berkowska, M. (2009). Singing proficiency in the majority: Normality and "phenotypes" of poor singing. *The neurosciences and music III—Disorders and plasticity: Annals of the New York Academy of Sciences, 1169,* 99–107. doi:10.1111/j.1749-6632.2009.04558.x

Dalla Bella, S., Giguère, J.-F., & Peretz, I. (2007). Singing proficiency in the general population. *Journal of the Acoustical Society of America, 121,* 1182–1189. doi:10.1121/1.2427111

Dalla Bella, S., Peretz, I., & Aronoff, N. (2003). Time course of melody recognition: A gating paradigm study. *Perception & Psychophysics, 65,* 1019–1028. doi:10.3758/BF03194831

Darwin, C. J., Turvey, M. T., & Crowder, R. G. (1972). An auditory analogue of the Sperling partial report procedure: Evidence for brief auditory storage. *Cognitive Psychology, 3,* 255–267.

Deutsch, D. (1969). Music recognition. *Psychological Review, 76,* 300–307.

Deutsch, D. (1970). Tones and numbers: Specificity of interference in short-term memory. *Science, 168,* 1604–1605.

Deutsch, D. (1972). Effect of repetition of standard and of comparison tones on recognition memory for pitch. *Journal of Experimental Psychology, 93,* 156–172. doi:10.1037/h0032496

Deutsch, D. (1973). Interference in memory between tones adjacent in the musical scale. *Journal of Experimental Psychology, 100,* 228–231.

Deutsch, D. (1974). Generality of interference by tonal stimuli in recognition memory for pitch. *Quarterly Journal of Experimental Psychology, 30,* 282–287.

Deutsch, D. (1975). The organization of short-term memory for a single acoustic attribute. In D. Deutsch & J. A. Deutsch (Eds.), *Short-term memory* (pp. 107–151). New York, NY: Academic Press.

Deutsch, D. (1980). The processing of structured and unstructured tonal sequences. *Perception & Psychophysics, 28*(5), 381–389. doi:10.3758/BF03204881

Deutsch, D. (1982). *The psychology of music* (1st ed.). New York, NY: Academic Press.

Deutsch, D. (1999). The processing of pitch combinations. In D. Deutsch (Ed.), *The psychology of music* (2nd ed., pp. 349–411). New York, NY: Academic Press.

DeWitt, L. A., & Crowder, R. G. (1986). Recognition of novel melodies after brief delays. *Music Perception, 3,* 259–274. doi:10.2307/40285336

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

Dowling, W. J. (1982). Melodic information processing and its development. In D. Deutsch (Ed.), *The psychology of music* (1st ed., pp. 413–429). New York, NY: Academic Press.

Dowling, W. J., & Bartlett, J. C. (1981). The importance of interval information in long-term memory for melodies. *Psychomusicology, 1,* 30–49. doi:10.1037/h0094275

Dowling, W. J., & Fujitani, D. S. (1971). Contour, interval, and pitch recognition in memory for melodies. *Journal of the Acoustical Society of America, 49,* 524–531. doi:10.1121/1.1912382

Free Software Foundation. (2007). *Gnu public license 3.0.* Retrieved 4 February, 2015, from www.gnu.org/copyleft/gpl.html

Frieler, K., Müllensiefen, D., Fischinger, T., Schlemmer, K., Jakubowski, K., & Lothwesen, K. (2013). Replication in music psychology. *Musicae Scientiae, 17,* 265–276. doi:10.1177/1029864913495404

Hallam, S., Cross, I., & Thaut, M. (2009). *The Oxford handbook of music psychology.* Oxford, UK: Oxford University Press.

Halpern, A. R. (1984). Perception of structure in novel music. *Memory & Cognition, 12,* 163–170. doi:10.3758/BF03198430

Halpern, A. R., & Bower, G. H. (1982). Musical expertise and melodic structure in memory for musical notation. *American Journal of Psychology, 95,* 31–50. doi:10.2307/1422658

Halpern, A. R., & Müllensiefen, D. (2008). Effects of timbre and tempo change on memory for music. *Quarterly Journal of Experimental Psychology, 61,* 1371–1384. doi:10.1080/17470210701508038

Halpern, A. R., & O'Connor, M. G. (2000). Implicit memory for music in Alzheimer's disease. *Neuropsychology, 14,* 391–397. doi:10.1037/0894-4105.14.3.391

Hébert, S., & Peretz, I. (1997). Recognition of music in long-term memory: Are melodic and temporal patterns equal partners? *Memory & Cognition, 25,* 518–533. doi:10.3758/BF03201127

Huron, D. (1994). Humdrum [Computer software]. Columbus, OH: Author. Retrieved from www.musiccog.ohio-state.edu/Humdrum/

Huron, D. (1996). The melodic arch in Western folksongs. *Computing in Musicology, 10,* 3–23.

Huron, D. (2006). *Sweet anticipation: Music and the psychology of expectation*. Cambridge, MA: MIT Press.

Hutchins, S., Larrouy-Maestri, P., & Peretz, I. (2014). Singing ability is rooted in vocal-motor control of pitch. *Attention, Perception, & Psychophysics, 76,* 2522–2530. doi:10.3758/s13414-014-0732-1

Jacoby, L. (1991). A process dissociation framework: Separating automatic from intentional uses of memory. *Journal of Memory and Language, 30,* 513–541. doi:10.1016/0749-596X(91)90025-F

Jakubowski, K., & Müllensiefen, D. (2013). The influence of music-elicited emotions and relative pitch on absolute pitch memory for familiar melodies. *Quarterly Journal of Experimental Psychology, 66,* 1259–1267. doi:10.1080/17470218.2013.803136

Jeffreys, H. (1961). *The theory of probability* (3rd ed.). Oxford, UK: Oxford University Press.

John, B. E., & Bass, L. (2001). Usability and software architecture. *Behaviour and Information Technology, 20,* 329–338. doi:10.1080/01449290110081686

Johnson-Laird, P. N., Kang, O. E., & Leong, Y. C. (2012). On musical dissonance. *Music Perception, 30,* 19–35. doi:10.1525/mp.2012.30.1.19

Krumhansl, C. (1979). The psychological representation of musical pitch in a tonal context. *Cognitive Psychology, 11,* 346–374. doi:10.1016/0010-0285(79)90016-1

Krumhansl, C. L. (1991). Music psychology: Tonal structures in perception and memory. *Annual Review of Psychology, 42,* 277–303. doi:10.1146/annurev.ps.42.020191.001425

Krumhansl, C. L., & Kessler, E. (1982). Tracing the dynamic changes in perceived tonal organization in a spatial representation of musical keys. *Psychological Review, 89,* 334–368. doi:10.1037/0033-295X.89.4.334

Levitin, D. (1994). Absolute memory for musical pitch: Evidence from the production of learned melodies. *Perception & Psychophysics, 56,* 414–423. doi:10.3758/BF03206733

Levitin, D., & Tirovolas, A. K. (2010). Music cognition and perception. In E. B. Goldstein (Ed.), *Sage encyclopedia of perception* (pp. 599–606). Thousand Oaks, CA: Sage.

Macaulay, C., Sloan, D., Jiang, X., Forbes, P., Loynton, S., Swedlow, J. R., & Gregor, P. (2009). Usability and user-centered design in scienfitic software development. *IEEE Software, 26,* 96–102. doi:10.1109/MS.2009.27

Mavromatis, P., & Farbood, M. M. (2012). The effect of tonal context on short-term memory for pitch. In E. Cambouropoulos, C. Tsougras, P. Mavromatis, & K. Pastiadis (Eds.), *Proceedings of the 12th International Conference on Music Perception and Cognition and the 8th Triennial Conference of the European Society for the Cognitive Sciences of Music* (pp. 674–675). Greece: Thessaloniki.

Mikutta, C. A., Maissen, G., Altorfer, A., Strik, W., & Koenig, T. (2014). Professional musicians listen differently to music. *Neuroscience, 268,* 102–111. doi:10.1016/j.neuroscience.2014.03.007

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review, 101,* 343–352. doi:10.1037/0033-295X.101.2.343

Müllensiefen, D. (2009a). FANTASTIC: Feature ANalysis Technology Accessing STatistics (in a corpus) [Computer software]. Retrieved from www.doc.gold.ac.uk/isms/m4s/FANTASTIC.zip

Müllensiefen, D. (2009b). *FANTASTIC: Feature ANalysis Technology Accessing STatistics (in a corpus)* (Technical Report v1.5). Retrieved from www.doc.gold.ac.uk/isms/m4s/FANTASTIC_docs.pdf

Müllensiefen, D., & Frieler, K. (2006). SIMILE [Computer software]. Retrieved 3 April, 2013, from www.mu-on.org

Müllensiefen, D., & Frieler, K. (2007). Modelling experts' notions of melodic similarity. *Musicae Scientiae, 11,* 183–210. doi:10.1177/102986490701100108

Müllensiefen, D., & Halpern, A. (2014). The role of features and context in recognition of novel melodies. *Music Perception, 31,* 418–455. doi:10.1525/MP.2014.31.5.418

Müllensiefen, D., & Pendzich, M. (2009). Court decisions on music plagiarism and the predictive value of similarity algorithms. *Musicæ Scientiæ, 13,* 257–295. doi:10.1177/102986490901300111

Müllensiefen, D., & Wiggins, G. A. (2011). Sloboda & Parker's recall paradigm for melodic memory: A new computational perspective. In I. Deliege & J. Davidson (Eds.), *Music and the mind: Essays in honour of John Sloboda* (pp. 161–188). Oxford, UK: Oxford University Press.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*, aac4716. doi:10.1126/science.aac4716

Peretz, I., Gaudreau, D., & Bonnel, A.-M. (1998). Exposure effects on music preference and recognition. *Memory & Cognition, 26,* 884–902. doi:10.3758/BF03201171

Rajaram, S. (1996). Perceptual effects on remembering: Recollective processes in picture recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22,* 365–377.

Schacter, D. L. (1987). Implicit memory: History and current status. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 13,* 501–518. doi:10.1037/0278-7393.13.3.501

Schacter, D. L., & Church, B. A. (1992). Auditory priming: Implicit and explicit memory for words and voices. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18,* 915–930. doi:10.1037/0278-7393.18.5.915

Schacter, D. L., & Wiseman, A. L. (2006). Reducing memory errors: The distinctiveness heuristic. In R. R. Hunt (Ed.), *Distinctiveness and memory* (pp. 89–107). New York, NY: Oxford University Press.

Schellenberg, E. G., & Habashi, P. (2015). Remembering the melody and timbre, forgetting the key and tempo. *Memory & Cognition, 43,* 1021–1031. doi:10.3758/s13421-015-0519-1

Schellenberg, E. G., Peretz, I., & Vieillard, S. (2008). Liking for happy- and sad-sounding music: Effects of exposure. *Cognition and Emotion, 22,* 218–237. doi:10.1080/02699930701350753

Schmuckler, M. A. (1997). Expectancy effects in memory for melodies. *Canadian Journal of Experimental Psychology, 51,* 292–306. doi:10.1037/1196-1961.51.4.292

Schulkind, M. D. (2004). Serial processing in melody identification and the organization of musical semantic memory. *Perception & Psychophysics, 66,* 1351–1362. doi:10.3758/BF03195003

Schulze, K., Zysset, S., Mueller, K., Friederici, A. D., & Koelsch, S. (2010). Neuroarchitecture of verbal and tonal working memory in nonmusicians and musicians. *Human Brain Mapping, 32,* 771–783. doi:10.1002/hbm.21060

Sloboda, J. A., & Parker, D. H. (1985). Immediate recall of melodies. In P. I. C. Howell & R. West (Eds.), *Musical structure and cognition* (pp. 143–167). London, UK: Academic Press.

Snyder, B. (2000). *Music and memory: an introduction*. Cambridge, MA: MIT Press.

Snyder, B. (2009). Memory for music. In S. Hallam, I. Cross, & M. Thaut (Eds.), *The Oxford handbook of music psychology* (pp. 107–117). Oxford, UK: Oxford University Press.

Sperling, G. (1960). The information available in brief visual presentations. *Psychological Monographs: General and Applied*, *74*(11, Whole No. 498), 1–29.

Stalinski, S. M., & Schellenberg, E. G. (2013). Listeners remember music they like. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 39,* 700–716. doi:10.1037/a0029671

Steinbeck, W. (1982). *Struktur und Ähnlichkeit: Methoden automatisierter Melodieanalyse*. Kassel, Germany: Bärenreiter.

Stevens, C., & Byron, T. (2009). Universals in music processing. In S. Hallam, I. Cross, & M. Thaut (Eds.), *The Oxford handbook of music psychology* (pp. 14–23). Oxford, UK: Oxford University Press.

Temperley, D. (2007). *Music and probability*. Cambridge, MA: MIT Press.

Tulving, E. (1985). Memory and consciousness. *Canadian Psychology, 26,* 1–12. doi:10.1037/h0080017

Vuvan, D. T., Podolak, O. M., & Schmuckler, M. A. (2014). Memory for musical tones: The impact of tonality and the creation of false memories. *Frontiers in Psychology, 5*(582), 1–18. doi:10.3389/fpsyg.2014.00582

Warker, J. A., & Halpern, A. R. (2005). Musical stem completion: Humming that note. *American Journal of Psychology, 118,* 567–585. doi:10.2307/30039086

White, B. W. (1960). Recognition of distorted melodies. *American Journal of Psychology, 73,* 100–107. doi:10.2307/1419120

Williamson, V., 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 & Cognition, 38,* 163–175. doi:10.3758/MC.38.2.163

Yonelinas, A. P. (2002). The nature of recollection and familiarity: A review of 30 years of research. *Journal of Memory and Language, 46,* 441–517. doi:10.1006/jmla.2002.2864

Zajonc, R. B. (1968). Attitudinal effects of mere exposure. *Journal of Personality and Social Psychology, 9,* 1–27. doi:10.1037/h0025848