Investigating the Recency Effect in Intonation Drift

Farrah Sa’adullah

Master of Research

The MARCS Institute for Brain, Behaviour and Development

Western Sydney University
The work presented in this thesis is, to the best of my knowledge and belief, original except as acknowledged in the text. I hereby declare that I have not submitted this material, either in full or in part, for a degree at this or any other institution.

Farrah Sa’adullah
# Table of Contents

Acknowledgements .................................................................................................................. 6

Abstract ..................................................................................................................................... 7

1. Introduction .......................................................................................................................... 8

1.1. Why Does Intonation Drift Occur? .................................................................................... 8

1.1.1. Physiological mechanisms. ............................................................................................ 9

1.1.2. Serial position effect. .................................................................................................... 9

1.1.3. Implications of the serial position effect. ..................................................................... 11

1.2. Methods and Results From Past Studies Finding the Just Noticeable Difference .... 12

1.2.1. Just noticeable frequency difference ............................................................................. 12

1.2.1.1. Discrimination tasks ................................................................................................. 12

1.2.1.1.1. Length of time between stimuli presentations ....................................................... 13

1.2.1.1.2. Stimuli length and intonation .............................................................................. 13

1.2.1.1.3. Timbre .................................................................................................................. 15

1.2.1.2. Staircases .................................................................................................................. 16

1.2.1.3. Production tasks ....................................................................................................... 18

1.2.2. Just noticeable interval difference. ................................................................................ 18

1.3. Summary ............................................................................................................................ 19

1.4. Hypotheses ........................................................................................................................ 20

2. Experiment 1: Determining the Just Noticeable Interval Difference of 12 Chromatic Intervals ......................................................................................................................... 23

2.1. Method .................................................................................................................................. 23

2.1.1. Design ............................................................................................................................. 23

2.1.2. Participants ..................................................................................................................... 26

2.1.3. Apparatus ....................................................................................................................... 27
2.1.4. Materials ........................................................................................................... 27
2.1.4.1. Intervals ........................................................................................................... 28
2.1.4.2. Musical experience questionnaire. ................................................................. 28
2.1.5. Procedure .......................................................................................................... 29
2.2. Results ................................................................................................................... 31
2.2.1. Descriptive Statistics ......................................................................................... 31
2.2.2. Inferential Statistics ........................................................................................... 35
2.3. Discussion .............................................................................................................. 41
2.3.1. Small Variation of Average Just Noticeable Interval Differences and Musical
       Training ...................................................................................................................... 41
2.3.2. Interval Sizes Influencing the Just Noticeable Interval Difference ................. 42
2.3.3. Absolute Detune Start Value Influencing the Just Noticeable Interval Difference
       ................................................................................................................................. 44
3.1. Method .................................................................................................................. 46
3.1.1. Design ............................................................................................................... 46
3.1.2. Participants ....................................................................................................... 47
3.1.3. Apparatus .......................................................................................................... 48
3.1.3.1. Slider ............................................................................................................. 48
3.1.3.2. Musical experience questionnaire. ................................................................. 49
3.1.4. Materials .......................................................................................................... 49
3.1.4.1. Single pitches ............................................................................................... 49
3.1.4.2. Musical experience questionnaire. ............................................................... 50
3.1.5. Procedure ........................................................................................................... 50
3.1.5.1. Task 1: Single pitch reproduction................................................................. 50
3.1.5.2. Task 2: Single pitch reproduction after a delay ........................................51

3.2. Results .............................................................................................................51

3.2.1. Descriptive Statistics ..............................................................................52

3.2.1.1. Task 1: Single pitch reproduction .......................................................52

3.2.1.2. Task 2: Single pitch reproduction after a delay ..................................52

3.2.2. Inferential Statistics ...............................................................................52

3.3. Discussion ....................................................................................................59

3.3.1. Time Decay Increases Pitch Error .........................................................60

3.3.2. Higher Frequencies of Pitch Decrease Pitch Error ..............................60

3.3.3. Effect of Gender on Pitch Error ..............................................................61

4. Experiment 3: Pitch Reproduction of the First Note of Detuned Melodies .......62

4.1. Method ...........................................................................................................62

4.1.1. Design .......................................................................................................62

4.1.2. Participants ...............................................................................................68

4.1.3. Apparatus .................................................................................................68

4.1.4. Materials ....................................................................................................69

4.1.4.1. Detuned melodies ..............................................................................69

4.1.4.2. Confidence ratings ............................................................................70

4.1.4.3. Musical experience questionnaire ....................................................70

4.1.5. Procedure ..................................................................................................70

4.2. Results ..........................................................................................................72

4.2.1. Descriptive Statistics .............................................................................73

4.2.2. Inferential Statistics ...............................................................................74

4.3. Discussion ....................................................................................................82
4.3.1. Pitch Error is Significantly Affected When Detune Size is Below the Just Noticeable Interval Difference ................................................................. 82

4.3.2. Pitch Error is Larger When Detuning Occurs ........................................ 84

5. General Discussion ...................................................................................... 85

5.1. Limitations and Future Studies ................................................................. 87

6. Conclusion .................................................................................................. 89

References ..................................................................................................... 91

Appendix A ................................................................................................... 98

Appendix B .................................................................................................. 108

Appendix C ................................................................................................. 110

Appendix D ................................................................................................. 111

Appendix E .................................................................................................. 112

Appendix F .................................................................................................. 113

Appendix G .................................................................................................. 114
Acknowledgements

Thank you to my supervisors Dr Jennifer MacRitchie and Dr Andrew Milne, I could not have done this thesis without you both. Thank you also to my parents and my friends.

Thank you Sandra, and thank you Mark.
Abstract

Intonation drift is a gradual drift out of tune by as much as a semitone that often occurs in unaccompanied choirs. This thesis explores intonation drift from the perspective of two serial position effect mechanisms: the recency interval mechanism and the primacy interval mechanism. The contribution of these two mechanisms to intonation drift is investigated over three experiments. In Experiment 1, it was assumed that intonation judgements of musical intervals in short-term memory would be made via comparison with reference intervals stored in long-term memory. As such, Experiment 1 determined the threshold value for when an interval was perceived to be out of tune. Experiment 2 investigated what might affect pitch error when reproducing a single pitch. Experiment 3 detuned melodies by various amounts as guided by the threshold value found in Experiment 1, in order to investigate if different detune amounts resulted in pitch error differences when reproducing the first note of a melody. Results indicated that the recency interval mechanism contributed to pitch error in addition to the primacy interval mechanism. Additionally, pitch errors in reproducing the first note of a melody appeared to plateau once the detuning was perceivable, and pitch errors were the largest when detuning was not perceivable. The latter finding may explain why pitch drift can be as much as a semitone in intonation drift. This research has implications for choir rehearsal strategies and vocal pedagogy.

Keywords: intonation drift, serial position effect, recency effect, primacy effect, just noticeable difference, just noticeable interval difference
1. Introduction

Singing is a form of vocal expression that transcends cultures and places (Dalla Bella, Giguère, & Peretz, 2007). It has been present in every known human culture since prehistoric times (Mauch, Frieler, & Dixon, 2014). Despite this, there has not been much research on the several different control processes involved in singing (Mauch et al., 2014). Perhaps the most important of these is the control of pitch, i.e. intonation (e.g. Watts, Barnes-Burroughs, Adrianopoulos, & Carr, 2003; Howard, 2007a), which is defined as singing and/or playing in tune (Mauch et al., 2014). In western music, this most commonly means being in tune with 12-tone equal temperament (12-TET) (Ferrer, 2007).

Intonation is central to many styles of singing (Mauch et al., 2014), and is the biggest factor for distinguishing between an untrained individual with singing talent and an untrained individual without singing talent (Watts et al., 2003). As such, staying in tune is a key focus of choir and a cappella (unaccompanied choir) rehearsals (Ganschow, 2013). Yet, unaccompanied choirs, in particular, report that they often pitch drift/slide in pitch/go out of tune gradually (Seaton, Sharp, & Pim, 2014). The perception of others going out of tune is reported as undesirable to listeners and performers (Watts et al., 2003). This gradual drift is known as intonation drift (ID; Seaton et al., 2014).

ID is a gradual change in pitch that most commonly occurs in unaccompanied singing (Seaton, et al., 2014). ID can occur in an upward direction, for example in solo folk singing (Müller, Grosche, & Wiering, 2010), but it is more frequently reported in a downward direction, e.g. in unaccompanied choirs (Mauch et al., 2014). In unaccompanied singing the amount of drift can be as much as semitone, which is the smallest interval (the distance between two notes) in western classical music (Seaton et al., 2014). The drift in ID can occur over tens of seconds or a whole piece (Seaton et al., 2014).

1.1. Why Does Intonation Drift Occur?
Despite ID being reported that it occurs often (Seaton et al., 2014), and despite choirs prioritising achieving accuracy of intonation (Ganschow, 2013), choirs also purportedly do not utilise any specific strategy to achieve more accurate intonation due to a lack of understanding of why ID occurs (Howard, 2007b). Therefore, the aim of this project is to investigate why ID occurs. In this section, previously discovered mechanisms will be reviewed, and towards the end of the section, two mechanisms (out of many that are possible) that will be the focus of this thesis will be discussed.

1.1.1. Physiological mechanisms. Previous research on why ID can occur in choirs has mainly explored issues in vocal production and physiological mechanisms (e.g. Mürbe, Pabst, Hofmann, & Sundberg, 2002; Mauch et al., 2014). For example, it has been found that poor-pitch singing (which refers to deficits in using pitch during singing; Welch, 1979) is a consequence of poor vocal-motor control (Pfordresher & Brown, 2007) and interference on auditory feedback (pitch feedback from the environment) and kinaesthetic feedback (muscle memory of the larynx) (Mürbe, Pabst, Hofmann, & Sundberg, 2002). A lack of auditory feedback was found to cause less accurate intonation in singers and kinaesthetic feedback was affected by different tempos and expressive techniques (Mürbe et al., 2002). There have also been observations of ID occurring because of a tendency towards singing in the just intonation tuning system (Howard, 2007b), poor concentration and tiredness (Seaton et al., 2014), and specific characteristics of the music (e.g. Dalla Bella et al., 2007; Seaton et al., 2014). However, the above mechanisms, unlike the next mechanism, do not take into account that ID occurs over time, over a series of notes (Seaton et al., 2014).

1.1.2. Serial position effect. One mechanism (out of many that are possible) that will be investigated in this thesis will here be called the recency interval mechanism (RIM), which is where the pitch/intonation of every note is judged with respect to that of the previous note. That is, each consecutive pair of notes is judged as an interval. This is based
on the serial position effect, which is where the last items in a series are recalled first and best (the recency effect; Murdock, 1962). This is a possible mechanism for why ID occurs because ID is often a very gradual drift, and small pitch changes may not be perceptible (e.g. Madsen & Geringer, 2004; Shepard & Jordan, 1984). For example, consider an 11-note melody that begins in tune on the first note, but, by the eleventh note, is out of tune by one semitone. This means that, given a linear detuning, the second to eleventh notes are progressively out of tune by 10 cents each (there are 100 cents in a semitone). That is, the interval (i.e. the distance between two pitches) made by the second and third notes, and the third and fourth, etc., are each widened or narrowed by 10 cents depending on the direction of the interval and the direction of detuning, i.e. whether the melody overall is going up or down by one semitone. The 10 cents error per interval in this example may be too small to be directly perceived. Since this is not recognised as an error, a singer is unlikely to correct their pitch, and these small and individually undetectable errors will potentially stack up over time to create a sizeable drift; that is, ID.

The other mechanism (out of many that are possible) that is assumed to be present will here be called the *primacy interval mechanism (PIM)*, which is where the first note/reference pitch is used to judge the intonation of every subsequent note heard as an interval to that first note. This is also based on the serial position effect, specifically the primacy effect, which is where the first items in the series are recalled second in time and second most accurately, after recalling the last items in the series (Murdock, 1962). PIM is assumed to be present because a reference pitch is normally heard before the rehearsal/performance of a piece by an unaccompanied choir begins, as there is an expectation that the singers will remember it and use it as a reference throughout the piece in order to stay in tune. However, the first note may be forgotten as a piece goes beyond the 6-
to 11-note storage limit of short-term memory (STM, which can hold a small amount of information for a short time) of pitch (Pembrook, 1987), therefore leading to ID.

The contribution of RIM and PIM to ID will be tested in the experiments in this thesis. To the best of my knowledge an investigation of the contribution of RIM and PIM to ID has not been done in previous research.

1.1.3. Implications of the serial position effect. In both RIM and PIM, one interval is judged for its intonation. In RIM, it is the interval between the two most recently heard notes. In PIM, it is the interval between the first note heard and the most recent note heard. The implication of both RIM and PIM is that both mechanisms are based on comparing the STM of an interval to the long-term memory (LTM, which holds material that has been previously learnt; Berz, 1995) of an interval that acts as a reference. For example, in a melody with the sequence of pitches C4, E4, G4, A4, G4, D4, F4, PIM compares the interval C4–F4 (which comprises recently heard pitches, so is stored in STM) with a ‘reference’ perfect 4th interval that is stored in LTM. RIM on the other hand compares the interval D4–F4 (which comprises the two most recently heard pitches, so is also stored in STM) with a ‘reference’ minor 3rd interval that is stored in LTM.

If a singer hears a discrepancy between their STM and LTM of an interval, that gives them the opportunity to take remedial action to correct their intonation. If they do not hear a discrepancy, then they do not have that opportunity to correct themselves. For example, consider a melody where the drift in tuning is slow enough such that each successive interval is out of tune by an amount that is below some supposed threshold, or the just noticeable interval difference (JNID). The result would be that for PIM, a singer would know that they are going out of tune and can take corrective action (because the tuning errors per interval would add up over time, meaning that eventually the drift would be substantial compared to the first note and therefore detectable). However, with RIM, they may not be able to take
corrective action because they are not aware that the interval is tuned incorrectly. This raises the question of how to methodologically investigate how different the STM interval needs to be from the LTM interval for them to be recognised as different. In other words, how the JNID would be measured. This is a novelty of this thesis as few studies have done this before.

Previous studies more commonly investigate the threshold value of perceiving a difference in two stimuli in STM. An example would be comparing two presented frequencies for whether they are the same or different. A comparison that involves two (or more) stimuli only in STM will be called in this thesis the just noticeable frequency difference (JNFD). Notwithstanding the discrepancy of a STM-STM comparison and a STM-LTM comparison, both are still trying to find the threshold value for perceiving a difference between the stimuli. Therefore, it is worth reviewing the methods and results of previous JNFD and JNID studies.

1.2. Methods and Results From Past Studies Finding the Just Noticeable Difference

There are many just noticeable difference studies in many areas of research, such as vision (e.g. Zhao, Chen, Zhu, Tan, & Yu, 2011), colour (e.g. Chou & Liu, 2008), and phonemes/speech perception (e.g. Huggins, 1972), as well as music perception. They share similar methods for determining the just noticeable difference. This section will focus on the methods that are used to find the just noticeable difference threshold value in JNFD (just noticeable frequency difference) and JNID (just noticeable interval difference) studies, and their results will be discussed.

1.2.1. Just noticeable frequency difference. This section reviews the three most common methods of finding the JNFD, and their results, which can vary widely depending on the method used and the stimuli presented.

1.2.1.1. Discrimination tasks. These tasks involve determining whether two (or more) presented stimuli are the same or different. The notes which have to be discriminated for
same/different will here be called \textit{comparison notes}. Features of the stimuli (e.g. the length of time between presentations of stimuli, the stimuli’s length and intonation, and the stimuli’s timbre) affect the JNFD.

The simplest version of the discrimination task involves discriminating between two comparison notes. For example, in a comparison of two notes varying between 2093 and 3750 ms in length, the JNFD for musicians was 20 cents (Hutchins & Peretz, 2012). The sections below review more complex JNFD discrimination tasks.

\textit{1.2.1.1.1. \textit{Length of time between stimuli presentations.}} When the length of time between two comparison notes for discrimination increases, the JNFD is affected. For example, a loss of pitch information occurs in STM when there is a delay before recall occurs (Williams, 1975). This is called time decay (Williams, 1975). The length of time before information is completely lost from STM is reportedly from approximately 20 seconds for letters and numbers (Peterson & Peterson, 1959). For music, some state that time decay in STM for pitch is similar to results that have been found in verbal auditory and visual STM research (Williams, 1975). For instance, minimal loss of information in STM was found for pitch discrimination of two comparison notes separated by 5 seconds of silence (Deutsch, 1970). However, others have found that time decay in STM for pitch operates quite differently. STM for music has been found to be 180 seconds (Kauffman & Carlsen, 1989), which is generally acknowledged to be not applicable to STM for non-musical information. Arguably, item decay, which is the presence of \textit{interfering notes} (i.e. notes that are in between/surrounding/close by to the comparison notes), causes more loss of information in STM for pitch (Williams, 1975).

\textit{1.2.1.1.2. \textit{Stimuli length and intonation.}} Determining the JNFD of longer presentations of stimuli is increasingly influenced by STM pitch capacity (Pembrook, 1987) because certain pitch information needs to be retained in order to make discrimination
judgements. For example, a scale discrimination ABX task demonstrated this, ABX being where stimulus A is presented, followed by stimulus B, and then stimulus X is the one that has to be judged for whether it is stimulus A or stimulus B. Stimulus A in this case was an ascending diatonic scale in 12-TET, and stimulus B was a manipulated 12-TET where the 3rd, 6th and 7th notes were sharpened or flattened by 6, 12, 18, 24, or 30 cents to equal degrees (Ward & Martin, 1961). It was found that the average threshold for getting the ABX task correct 75% of the time was approximately 20 cents (Ward & Martin, 1961). When the ABX task was shortened to an AX same/different discrimination task (12-TET and manipulated 12-TET), the threshold for answering the AX task correctly 75% of the time was approximately 10 cents (Ward & Martin, 1961). STM pitch capacity appears to be the limiting factor here as participants in this study reported that the discrimination would be easier if two scales rather than three were used (Ward & Martin, 1961). This concurs with loss of pitch information occurring more rapidly when there are more interfering notes (Deutsch, 1972).

The JNFD of two comparison notes can also become smaller if there are interfering notes that have accurate intonation, compared to two standalone comparison notes. For example, Warrier and Zatorre (2002) compared the discrimination of two comparison notes (isolated condition) versus two notes preceded by a completely in tune melody (melodic condition). In both conditions, the first note was always in tune and the second note was either 0, 17, 35 or 52 cents higher than the first note (Warrier & Zatorre, 2002). It was found in the isolated condition that there were no significant differences across the different amounts of pitch deviations for discriminating whether the second note was higher than the first (Warrier & Zatorre, 2002). In the melodic condition however, the bigger pitch deviations of 35 and 52 cents were significantly more detectable (Warrier & Zatorre, 2002).

A similar result was found in another discrimination task containing an F#4 of 30 seconds duration going out of tune over the last 25 seconds at 2 cents per second ascending or
descending, such that the maximum frequency deviation was ±50 cents (Madsen, Edmonson, & Madsen, 1969). The change in frequency of the note was best perceived during the first 5 seconds it started occurring, i.e. the first 10 cents of pitch change (Madsen et al., 1969), as it is in closest proximity to the stable in-tune pitch.

Conversely, if the intonation of the interfering notes is inaccurate, it decreases the ability to discriminate between comparison notes (i.e. results in a bigger JNFD). This was demonstrated when each interval of a C major ascending diatonic scale starting on middle C (C4) and spanning an octave was stretched by a factor of 13/12, such that by the last note the scale had reached C#5 (i.e. one semitone higher than the last note should have been had the scale not been stretched; Shepard & Jordan, 1984). After first listening to the stretched scale, participants were then played either a C4 or a C#4 and 86% of them rated C4 as lower than the first note heard, while 63% rated C#4 the same as the first note (Shepard & Jordan, 1984). Therefore, these results overall suggest that pitch discrimination is less accurate and a larger JNFD results when interfering notes have inaccurate intonation, and the JNFD is smaller when interfering notes have accurate intonation. Also, the JNFD is smaller when the stimuli comparisons do not exceed the STM pitch capacity.

1.2.1.3. Timbre. The JNFD can be greatly influenced by the timbre of the stimuli being compared. To see what would affect it would be relevant for deciding what sound to use in the experiments in this thesis. Tones with the presence of harmonics (i.e. not just the fundamental frequency) appear to decrease the JNFD of two comparison notes, because changes can be detected in the harmonics as well as in the fundamental (Zeitlin, 1964). For example, a complex tone compared to a pure tone made it easier to judge that the middle of three tones was higher/lower than the two identical first and third tones (Zeitlin, 1964). Tones that have a stable pitch within a timbre also result in a smaller JNFD, compared to tones that have a less stable pitch within a timbre (Hutchins & Peretz, 2012). For example, a violin
timbre allowed for better discrimination of two notes than a voice timbre (Hutchins, Roquet, & Peretz, 2012), and for musicians discriminating between two comparison notes, a synthesised vocal tone resulted in a lower JNFD of 20 cents compared to 30 cents for a natural voice timbre (Hutchins & Peretz, 2012). Complex tones that are stable within their timbre may therefore allow for more accuracy in noticing differences in pitch, or a smaller JNFD.

1.2.1.2. Staircases. Another method of determining the JNFD is by using a staircase, which is an adaptive procedure that can determine the average value of sensory thresholds (Linschoten, Harvey, Eller, & Jafek, 2001) and is commonly used for vision/colour differences (e.g. Chou & Liu, 2008) and in psychoacoustic testing (Levitt, 1971). It is a forced choice procedure, meaning that the participant has to answer ‘yes’ or ‘no’ for whether they detect a difference or not. Staircases for determining the JNFD involve the discrimination of two comparison notes, normally starting with quite an easily detectable difference between them. The difference is then reduced in steps (the staircase steps down) until the participant makes a mistake (by reporting that there is no difference, when there is), which causes the staircase to step up (make the difference more detectable again). The point at which a staircase changes from stepping down to stepping up (or vice versa) is called a reversal, or a turnaround. The number of turnarounds and the size of the steps a staircase will have are based on the level of accuracy required for determining the threshold of detecting a difference and the parameters that will be estimated (García-Pérez, 1998; Levitt, 1971). The threshold is determined by averaging the values of the turnarounds, although sometimes turnaround values are discarded if they are from the initial phase, which is when the rules of the staircase are often modified, e.g. step sizes may be different (Kaernbach, 1991).

There are many different methods of stepping through a staircase. The size of the steps downward can differ from the size of the steps upwards, which is reportedly a stable
and consistent method (Kaernbach, 1991). Normally though, staircases differ on how many correct answers must be given consecutively before a staircase steps down (García-Pérez, 1998). For example, a 2 down 1 up method requires a correct detection of a difference twice before the staircase steps down, and only one incorrect detection for the staircase to step up. An increasing number of consecutive correct detections before a staircase steps down leads to a better chance of the threshold being correctly identified (García-Pérez, 1998). A 1 down 1 up method has a 77.85% chance of detecting the threshold correctly (García-Pérez, 1998) and gives the 50% correct threshold (Brown, 1996), while a 4 down 1 up method has a 85.84% chance of detecting the threshold correctly (García-Pérez, 1998).

Staircases in sensory threshold studies (on both psychoacoustics and vision) are usually only STM-STM comparisons. For example, a 1 down 1 up staircase method was used to determine the average threshold difference before two notes could be discriminated (Tervaniemi, Just, Koelsch, Widmann, & Schröger, 2005). Steps downward decreased the distance between the two notes by 1 Hz, and steps upwards increased the distance between the two notes by 3 Hz (Tervaniemi et al., 2005). One note was always 528 Hz, and the other was between 529 and 558 Hz (Tervaniemi et al., 2005). The average JNFD found was 3.8 Hz, which is approximately 12.4 cents (Tervaniemi et al., 2005). In another study, a 2 down 1 up staircase method was used for an interval discrimination task, where two intervals were presented for discrimination (McDermott, Keebler, Micheyl, & Oxenham, 2010). The intervals had different first notes, and the step sizes first increased or decreased by a factor of 4, then 2, then the square root of 2 (McDermott et al., 2010). Musicians in this study were found to have an interval difference threshold of below a semitone (McDermott et al., 2010).

Since staircases are normally only used in STM-STM comparisons, the staircase method will therefore be adapted for STM-LTM comparisons of an interval in order to gain a value for the JNID (see Section 2.1.1.). The LTM of an interval will be assumed, as this is a
likely scenario in trained musicians (Krumhansl, 1991), and participants will be asked if the interval is out of tune.

1.2.1.3. Production tasks. These tasks involve reproducing a pitch on a device such as a slider, which has been used to determine the JNFD. A slider produces a seemingly continuous and unquantised output of pitch when it is touched, and has been used to reproduce pitches since it results in more accurate pitch reproduction than when using one’s voice (Hutchins & Peretz, 2012). The slider in Hutchins and Peretz’s (2012) study was used for a continuous/repeated comparison of two notes successively (never simultaneously/harmonically), and the final difference between the two notes was the JNFD. The note to be reproduced was continuously heard, except for when the slider was touched (Hutchins & Peretz, 2012). The JNFD in this scenario was found to be on average within 2 cents for musicians (Hutchins & Peretz, 2012).

1.2.2. Just noticeable interval difference. The most common method of determining the value of the JNID (which is of isolated musical intervals/stimuli in STM compared to a reference in LTM) is through adjustment methods (Burns, 1999). Adjustment methods usually investigate the perception of isolated octaves (Burns, 1999). They are used less frequently for other intervals (e.g. Rakowski, 1990). In an experiment using the adjustment method, normally a pair of successive tones are presented, one whose frequency is fixed, and the other (the variable tone) whose frequency can be adjusted by the participant (Burns, 1999). The participant is instructed to adjust the variable tone so that its relationship to the fixed tone is some specified musical interval (Burns, 1999). Examples of devices that are used for adjustment are tuning dials and variable-tone oscillators (Geringer, 1978; Rakowski, 1990), which are similar to the slider.

Of the few studies that used adjustment methods to investigate the JNID, one study used a tuning dial with a range of ±3 semitones so that participants could retune their own
performances of the ascending Mixolydian scale (a major scale starting on the fifth note) (Geringer, 1978). Participants had 4 seconds to retune each note of their performance, and it was found that all intervals tended to be tuned sharp (Geringer, 1978). Another study used a variable tone oscillator where participants were given the opportunity to adjust the 12 within-octave chromatic intervals, all of them being presented melodically (Rakośki, 1990). The two notes of each interval alternated continuously 10 times: the first note was 0.5 seconds in length, which was then followed by a 0.5 second break, followed by the second note of the interval that was also 0.5 seconds in length, followed by a 1.5 second break before the cycle repeated (Rakośki, 1990). It was found that small intervals tended to be tuned smaller and large intervals tended to be tuned larger than 12-TET intervals should be (Rakośki, 1990).

1.3. Summary

Accuracy of intonation is one of the main focal points for judging singing ability (e.g. Watts et al., 2003). Achieving intonation accuracy is thus a priority in choirs and a key focus of rehearsals. However, despite the importance of intonation and the practice time dedicated to improving it, ID still occurs.

RIM and PIM are the two proposed mechanisms out of many possible mechanisms for how accuracy of intonation is achieved. The contribution of RIM and PIM to ID will be investigated in this thesis. Both of them are based on the serial position effect. The first one, RIM, is based on the recency effect, where the most recent items in a series are recalled first and best (Murdock, 1962). Hence, in RIM, the intonation of the current note is judged as an interval to the previous note. ID may then occur if the error in tuning of the interval is below the threshold of perception, so corrective action cannot be taken. These errors subsequently accumulate over time, resulting in ID.

The second mechanism is PIM and is based on the primacy effect, which is where the first items in a series are recalled after the most recent items are recalled first and best.
(Murdock, 1962). PIM relates to the scenario where the first/reference pitch normally given to an unaccompanied choir is remembered and used to judge the intonation of the most recent note as an interval to the first note. However, the first note may be forgotten due to the STM pitch capacity (Pembrook, 1987), which would result in worse accuracy in judging the interval formed by the first note and the most recent note, and lead to the occurrence of ID.

In both RIM and PIM, a comparison of an interval in STM and an interval in LTM is required to judge the intonation of the interval in STM. Since an interval must be previously learnt, and previously learnt material is stored in LTM (Berz, 1995), the experiments in this thesis will investigate musicians only. The threshold of detecting a difference between the interval in STM and the interval in LTM is termed the JNID. Therefore it may be that ID occurs because the drift out of tune occurs below the JNID.

The research question for the thesis is: what contribution do RIM and PIM have to the occurrence of ID?

1.4. Hypotheses

To investigate this research question, there will be three experiments.

Experiment 1 is an exploratory study that will explore what the range of participants’ average JNIDs look like, and what the average threshold is for when an interval crosses over from being perceived to be in tune, to out of tune, when an interval in STM is being compared to an interval in LTM. This experiment will relate to both RIM and PIM, since both involve STM-LTM discriminations of an interval. The average JNID will be determined for all 12 intervals within an octave using a staircase procedure, where participants will be asked if the interval presented to them is out of tune. All 12 intervals will be included because certain intervals may affect the JNID more than others. A more comprehensive exploration of the JNID for standalone 12-TET intervals may be relevant for a situation in which ID occurs.

As Experiment 1 is exploratory, its aim is to:
• Determine the coefficient of variation (which is the standard deviation divided by the mean) of the absolute value of the average JNIDs from each participant;

Experiment 2 will investigate STM for a single pitch over time. Specifically, this experiment will investigate if time decay/silence affects pitch reproduction of a single pitch in STM, using a slider to reproduce the pitches. Investigating time decay will relate to how the STM pitch capacity may be affected without the presence of interfering notes. It will also provide a baseline for Experiment 3 (see next paragraph) since in Experiment 2 only time decay/silence in STM will be investigated (the single pitch plus the length of silence will be the length of a melody in Experiment 3), whereas in Experiment 3 there will be interfering notes in place of the silence potentially affecting STM for pitch. Additionally, due to past research on certain ranges of frequencies being easier to discriminate (e.g. Zeitlin, 1964; Moore, 1973), the frequencies of pitch used in this experiment will be explored for whether they affect STM for pitch.

The hypothesis for Experiment 2 is:

• Pitch reproduction of a single pitch will result in greater pitch error when there is a delay (silence/gap) before pitch reproduction, compared to when there is no delay/gap before pitch reproduction of a single pitch.

Experiment 3 will simulate ID by gradually detuning melodies up or down, with instructions to remember and reproduce on a slider the first note heard after the melody is over. The amount of detuning per note (otherwise called the detune size) will be guided by the JNID value found in Experiment 1. Hence, some of the melodies will be detuned at a rate where the detune size of selected consecutive notes is below the value of the JNID, some notes of the melody will have a detune size that is exactly the value of the JNID, and some will have a detune size that is above the value of the JNID. There will also be a control/baseline condition with no detuning. To ensure that the last note of every melody will
be out of tune by the same amount, the onset of detuning will occur at different points in the melody. Hence, detuning would start earlier in a melody with a smaller detune size, and detuning would start later in a melody with a greater detune size. Having different detune sizes as guided by the JNID is to investigate the contribution of RIM and PIM to ID, because if the different detune sizes affect pitch error when reproducing the first note of the melody, this would suggest that RIM and PIM both contribute to ID. However, if the different detune sizes do not affect pitch error when reproducing the first note of the melody (since all melodies would be detuned by the same amount by their last note), this would suggest that RIM does not contribute to ID.

The hypotheses for Experiment 3 are:

- The different detune sizes/rates of detuning will affect the magnitude of pitch error to different extents when reproducing the first note of the melody. This would indicate an effect of RIM. If the rate of detuning is below the JNID/threshold of perception, the resulting pitch error will be larger. If the rate of detuning is at the value of the JNID, the pitch error will be smaller than the pitch error from the rate of detuning that is below the JNID. If the rate of detuning is above the value of the JNID, the pitch error will be smaller than the pitch error from the rate of detuning that is at the value of the JNID. If these differences in pitch error do not occur, this would suggest that RIM does not contribute to ID;

- All detune sizes would result in a larger pitch error when reproducing the first note of the melody, compared to when there is no detuning. This would be an effect of RIM contributing to pitch error in addition to PIM, where if there is no detuning, pitch reproduction of the first note of a melody would be facilitated by in-tune notes, and if there is detuning, pitch reproduction of the first note of a melody would be hindered by out-of-tune notes.
In Experiment 2 and 3, reproducing the pitch of a particular note will be performed on a slider due to limitations when using the voice. For example, poor pitch singers have worse vocal pitch production skill, which may mask perceptual ability, as poor pitch singers have been found to not differ from good singers in terms of pitch discrimination accuracy (Pfordresher & Brown, 2007). The use of a slider would standardise production skill across participants, taking out the vocal production factor so that RIM and PIM’s contributions to ID can be investigated without a potential confounding variable.

2. Experiment 1: Determining the Just Noticeable Interval Difference of 12 Chromatic Intervals

The aim of this experiment was to find the JNID of 12 chromatic intervals (from semitone to octave) using an adaptive staircase procedure and, if possible, take the average of them and use the resulting value as a guide for how much the melodies in Experiment 3 should be detuned. The unison interval was not included because this has been covered before in JNFD experiments (a STM-STM comparison).

2.1. Method

2.1.1. Design

The independent variables were:

- **Interval Size** (12 levels: minor 2\textsuperscript{nd}, major 2\textsuperscript{nd}, minor 3\textsuperscript{rd}, etc., to an octave (coded as 1 to 12)), a categorical variable;

- **Interval Direction** (2 levels: ascending (second note of the interval being higher than the first note, coded as ‘1’), descending (second note of the interval being lower than the first note, coded as ‘–1’)), a categorical variable;

- **Absolute Detune Start Value**, a categorical variable which is the absolute value of the staircase’s start value for an interval’s second note (2 levels: 0 cents (second
note of the interval started in tune, coded as ‘0’), 50 cents (second note started 50 cents out of tune, coded as ‘1’));

- **Staircase Direction**, a categorical variable which specifies what amount of pitch deviation the second note of the interval should take for the next step of the staircase when the participant correctly identifies the interval as out of tune or in tune (2 levels: +10 cents (coded as ‘1’), –10 cents (coded as ‘−1’)). **Staircase Direction** also characterises whether **Absolute Detune Start Value** is +50 cents or –50 cents. A −50 cents detune would always be a **Staircase Direction** +10 cents and a 50 cents detune would always be a **Staircase Direction** –10 cents. A 0 cents **Absolute Detune Start Value** would be specified by **Staircase Direction** as either +10 cents or −10 cents.

All independent variables were within-subjects except for **Interval Size**, which was a rolling design where each participant received 16 pre-decided different staircase conditions (a staircase condition here refers to the values the staircase is assigned to with respect to **Interval Size**, **Interval Direction**, **Absolute Detune Start Value**, and **Staircase Direction**; see Table 1). A rolling design is where, for example, the first participant receives staircase conditions numbers 1 to 16, the second participant receives staircase conditions numbers 5 to 20, the third participant receives staircase conditions numbers 9 to 24, and so on.

Only the second note of the interval was manipulated according to **Absolute Detune Start Value** because it had the most relevance to Experiment 3 (where the first note of every melody would always be in tune with 12-TET).

The dependent variable was the JNID measured in cents. To determine this, a 1 down 1 up staircase procedure using equal step sizes and with 8 turnarounds was used, because 1 down 1 up converges on the stimulus level that has 50% positive responses (Brown, 1996).
Eight turnarounds were used because 6 to 8 turnarounds is the recommended procedure for 1 down 1 up staircases (Wetherill & Levitt, 1965).

A 1 down 1 up staircase is where, for example, the detuning of the second note of the interval starts at 50 cents out of tune. If the tuning error of the interval is detected correctly (i.e. the participant judges it to be ‘out of tune’), the staircase steps down to 40 cents, then 30 cents, etc., until the difference is no longer detected correctly. When this happens, the staircase reverses/steps back up by 10 cents (turnaround no. 1) until the difference is correctly detected again. When this happens, the staircase steps down again (turnaround no. 2).

A limit was put on how much the interval could go out of tune; the second note of an interval could only be up to 50 cents out of tune. This is because, beyond that, the interval may be confused with the next interval wider or narrower than the interval in question.

If a participant reached the ceiling limits of how much an interval could go out of tune (the ceiling limits being 0 cents and ±50 cents) and they answered incorrectly 4 times in a row for whether or not an interval was out of tune (i.e. responding “yes” it is out of tune when it was not, and “no” when it was), then the staircase ended after the fourth consecutive incorrect response and its data was not included in the analysis. Four consecutive incorrect responses was deemed to be a large enough buffer to prevent participants from accidentally ending a staircase early, in case for example they unintentionally gave an incorrect response.

Since staircases could end early (i.e. not reach 8 turnarounds), participants each heard 32 staircases in the experiment: every staircase was heard twice in order to have a better chance of getting data for every staircase. As a result, each staircase condition was tested 3 to 10 times across all participants ($M = 5.84, SD = 1.47$).

For a summary of what each of the staircases was like in terms of the independent variables for any one participant, see Table 1.
Table 1

*The 16 Different Staircase Conditions Received by a Participant*

<table>
<thead>
<tr>
<th>Staircase No.</th>
<th>Interval Size (semitones)</th>
<th>Interval Direction</th>
<th>Absolute Detune Start Value</th>
<th>Staircase Direction</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>#</td>
<td>##</td>
<td>###</td>
<td>######</td>
</tr>
<tr>
<td>2</td>
<td>#</td>
<td>##</td>
<td>###</td>
<td>######</td>
</tr>
<tr>
<td>3</td>
<td>#</td>
<td>##</td>
<td>###</td>
<td>######</td>
</tr>
<tr>
<td>4</td>
<td>#</td>
<td>##</td>
<td>###</td>
<td>######</td>
</tr>
<tr>
<td>5</td>
<td>#</td>
<td>##</td>
<td>###</td>
<td>######</td>
</tr>
<tr>
<td>6</td>
<td>#</td>
<td>##</td>
<td>###</td>
<td>######</td>
</tr>
<tr>
<td>7</td>
<td>#</td>
<td>##</td>
<td>###</td>
<td>######</td>
</tr>
<tr>
<td>8</td>
<td>#</td>
<td>##</td>
<td>###</td>
<td>######</td>
</tr>
<tr>
<td>9</td>
<td>#</td>
<td>##</td>
<td>###</td>
<td>######</td>
</tr>
<tr>
<td>10</td>
<td>#</td>
<td>##</td>
<td>###</td>
<td>######</td>
</tr>
<tr>
<td>11</td>
<td>#</td>
<td>##</td>
<td>###</td>
<td>######</td>
</tr>
<tr>
<td>12</td>
<td>#</td>
<td>##</td>
<td>###</td>
<td>######</td>
</tr>
<tr>
<td>13</td>
<td>#</td>
<td>##</td>
<td>###</td>
<td>######</td>
</tr>
<tr>
<td>14</td>
<td>#</td>
<td>##</td>
<td>###</td>
<td>######</td>
</tr>
<tr>
<td>15</td>
<td>#</td>
<td>##</td>
<td>###</td>
<td>######</td>
</tr>
<tr>
<td>16</td>
<td>#</td>
<td>##</td>
<td>###</td>
<td>######</td>
</tr>
</tbody>
</table>

*Note.* # any number from 1–12, ## either ascending or descending, ### either 0 cents or 50 cents, ###### either +10 cents or –10 cents. All values remain the same for each staircase once the experiment starts till it ends for each participant.

### 2.1.2. Participants

Participants were 28 musicians (19 females, 9 males) aged 18 to 45 years old ($M = 27.8$ years, $SD = 7.2$), with 5 or more years of formal music training. Four participants had 5 years of formal training, 9 participants had 6 to 9 years of formal training, and 15 participants had 10 or more years of formal training. These categories are according to the musical
experience questionnaire they completed (see Section 2.1.4.2.). Participants were recruited via social media, flyers put up on Western Sydney University (WSU) campuses, and from the MARCS Institute for Brain, Behaviour and Development Music Cognition and Action participant database. All participants resided in Sydney and were reimbursed $20/hr to cover travel expenses.

The eligibility criteria were, (a) aged 18 to 45 years old, and (b) musicians with 5 or more years of formal music training, which was defined as private music lessons on a musical instrument or voice (Hutchins & Peretz, 2012). The first criterion was selected due to the prevalence of hearing loss increases as age increases, such that at 50 to 65 years of age, 25% of this age group reportedly have at least mild hearing loss (Liu & Yan, 2007). The upper limit of the age criterion is therefore a conservative choice in order to be fairly sure that hearing loss is not introduced as an uncontrolled variable. The second criterion was chosen because pitch discrimination has been found to be more accurate amongst musicians with more experience and age (Madsen et al., 1969), and because categorical perception of intervals appears to be proportional to musical training while non-musicians were found to exhibit no categorical boundaries for intervals (Krumhansl, 1991).

2.1.3. Apparatus

The experiment took place with participants individually using a Mac laptop with a computer mouse to progress through the experiment presented on the laptop. All auditory stimuli were presented in mono and heard binaurally through closed back over the ear headphones. The experiment was constructed, run on, and responses recorded by Max/MSP (Cycling ’74, San Francisco, CA), which is a program that can create audio, video and graphics for user interaction. The musical experience questionnaire (see Section 2.1.4.2.) was hosted on Qualtrics, an online data collection tool.

2.1.4. Materials
2.1.4.1. Intervals. The tones used to play the intervals in the stimuli were sawtooth waves put through a lowpass 3\textsuperscript{rd} order (\(-18\text{dB per octave}\)) Butterworth filter, with a cut-off frequency of 2756.25 Hz. This complex tone was used because it does not sound like any specific instrument due to the lack of characteristic temporal envelopes, but its steady-state spectrum is similar to what is produced by many common instruments such as strings, brass, and voice. A complex tone was also used as opposed to a pure tone because frequency changes are detected better in complex tones than in pure tones (Zeitlin, 1964), and because all musical instruments produce complex tones. Not using the sound of a particular instrument however was to avoid any bias from certain participants who may be more accustomed to hearing that sound (e.g. it may be their main instrument), as being familiar with a sound has been found to improve the accuracy of pitch matching with that sound (Ely, 1992).

The pitches used for the intervals were semitone steps from Bb\textsubscript{4} to G\textsubscript{6} inclusive, that is, 22 notes in total. This range was used because it is the same range as for all the melodies that would be used in Experiment 3. Hence the JNID found in Experiment 1 would translate directly to Experiment 3.

The first note of the interval was randomly chosen for each participant from the Bb\textsubscript{4} to G\textsubscript{6} range, but with additional constraints depending on what interval they were about to hear. For example, if they were about to hear an octave going upwards in direction (second note higher than the first), the first note could not be G\textsubscript{6} because it would take them outside the Bb\textsubscript{4} to G\textsubscript{6} range.

Each note of the interval was 1 second long, and there was no gap between the two notes of the interval.

2.1.4.2. Musical experience questionnaire. The Goldsmiths Musical Sophistication Index (Gold-MSI; Müllensiefen, Gingras, Musil, & Stewart, 2014) is a self-report
questionnaire and was used to determine participants’ individual differences in musical experience and expertise (i.e. musical skills and behaviours), as more experience may result in more accurate detection of pitch errors (Madsen et al., 1969). The Gold-MSI contains 5 factors: Active Engagement, Perceptual Abilities, Musical Training, Singing Abilities, and Emotions. A description of each of these can be found in Appendix A, as well as the questionnaire used, the items in each subscale, and the mean, standard deviation and range of possible scores per subscale.

The Gold-MSI also collected data on participants’ age and gender.

Data was also collected on whether or not participants had absolute pitch, which is the ability to identify or recreate a given musical note without the benefit of a reference tone. They could only answer either ‘yes’ or ‘no’ for whether or not they had absolute pitch.

2.1.5. Procedure

The experiment consisted of two computer tasks: the musical intervals experiment (completed on Max/MSP using the mouse), and a self-report questionnaire. All participants completed these two tasks in the same order.

Before the experiment began, participants were not explicitly told that 12-TET was the correct tuning to make their intonation judgements about intervals. Participants were also not told by how much the intervals were out of tune, only that the intervals may or may not be out of tune. Lastly, they were instructed not to sing, hum, or articulate in any way throughout the experiment, as doing this may affect their ability to judge whether an interval is in or out of tune.

The 16 staircases as seen in Table 1 had their values for the 4 independent variables pre-loaded prior to the participant starting the experiment. A first note for the interval of each staircase was randomly chosen and also pre-loaded prior to the participant starting the experiment. Additionally, each step up or down for each staircase was presented in a random
order, which is a technique developed at the MARCS Institute for Brain, Behaviour and Development. For example, a step would be heard from staircase no. 3, then a step from staircase no. 15, then a step from staircase no. 11 (see Table 1 for the list of 16 staircases) etc., until every staircase was completed twice. This was to reduce participants’ potentially increasing sensitivity to the small tuning errors in intervals, and to reduce the likelihood of participants figuring out how the staircases worked. Each time a staircase was completed, the first note of the interval was randomly chosen again for the second iteration of the staircase.

Each step of each staircase was accompanied with the same question: “Is this [name of interval] interval ([note name] to [note name]) out of tune?”; for example, “Is this major 2\textsuperscript{nd} interval (C to D) out of tune?” or “Is this tritone (G to C#) out of tune?” The name of the interval and the note names were both given in case some participants were not familiar with standard interval naming conventions. Participants either clicked “Yes” or “No” buttons, using the mouse, to answer the question and progress through the staircase. The experiment thus proceeded in a way that is displayed in Figure 1.

![Figure 1](image_url)

**Figure 1.** The procedure of each step of any staircase.

Participants answered approximately 450 questions over 32 staircases. The number of questions each participant received depended on their answers (e.g. their consistency at judging whether an interval was out of tune or not). All participants were informed that the length of the musical intervals task may vary according to how they answered the questions
and how long they took to answer each question. To avoid participant fatigue, participants were encouraged to keep moving through the task at a steady pace and not think about their answer too much if they found themselves uncertain or stuck on a question. They were all informed that they could take a break from the experiment after a question/one step of a staircase was completed, or when they reached the break screen, which occurred every 100 questions (which meant it was possible that a break would occur during the progress through multiple staircases). Participants were able to track their progress through these 100 questions with a progress bar at the top of the screen, and after each 100 questions were complete, they received an onscreen joke about musicians.

After participants finished the musical intervals task, they completed the Gold-MSI (Müllensiefen et al., 2014), which included questions on demographics.

2.2. Results

2.2.1. Descriptive Statistics

The aim of this experiment was to determine the coefficient of variation of the average JNIDs from each participant. The coefficient of variation is calculated by dividing the standard deviation of the average JNIDs from each participant by the average of the average JNIDs from each participant.

The average JNID from each participant was averaged over all the staircases that they reached 8 turnarounds for. If 8 turnarounds were not reached, that staircase trial was rejected and not included in the participant’s final average JNID, nor the overall JNID of all participants.

The value at each turnaround in a staircase was averaged except for the first one, because the first turnaround was part of the initial phase of the staircase and it is common practice to not include turnarounds from the initial phase where rules and/or step sizes are often modified (Kaernbach, 1991). In this experiment, the rule about the second note of the
interval not going beyond 0 or 50 cents out of tune means that in the initial phase, if a participant answers incorrectly 2 to 3 times at the start of a staircase, the first turnaround (when they finally answer correctly) would be the minimum/maximum out-of-tune value of the staircase, which would not make sense. Hence, the first turnaround was not included in the average for determining the JNID of a staircase, and only the last 7 turnaround values were averaged.

The absolute value of the JNID of each staircase was taken because some staircase conditions would have resulted in negative JNID values. For example, if the staircase started at −50 cents out of tune, the steps would all have been negative values (−50 cents, −40 cents, etc.). Consequently, the average JNID from each participant ranged from 13.7 to 32.5 cents ($M = 25.30$ cents, $SD = 4.72$ cents). The number of (complete) staircases for each participant ranged from 10 to 32 ($M = 21.04$, $SD = 6.06$). The coefficient of variation was 0.19, which indicated low variance in the distribution/19% variability. A histogram of the average JNIDs from each participant can be seen in Figure 2. All plots were made on Matlab R2016b.
Figure 2. The frequency of average JNIDs from all participants, with 6 bins according to Sturges’ formula (Sturges, 1926).

Across all participants, 561 staircases were completed. The coefficient of variation of the complete staircases across all participants was 0.45, meaning that there was moderate variance in this distribution/45% variability. A histogram of the absolute value of the JNIDs from each complete staircase trial (across all participants) can be seen in Figure 3.
Figure 3. The frequency of the absolute value of the JNIDs from each complete staircase, with 11 bins according to Sturges’ formula (Sturges, 1926).

The participant scores for the 5 Gold-MSI factors can be seen in Table 2.
Table 2

Descriptive Statistics For Participants' Gold-MSI Scores

<table>
<thead>
<tr>
<th>Gold-MSI Factor</th>
<th>Minimum</th>
<th>Maximum</th>
<th>M</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Active Engagement</td>
<td>27</td>
<td>59</td>
<td>44.14</td>
<td>8.29</td>
</tr>
<tr>
<td>Perceptual Abilities</td>
<td>41</td>
<td>64</td>
<td>52.53</td>
<td>6.45</td>
</tr>
<tr>
<td>Musical Training</td>
<td>28</td>
<td>46</td>
<td>39.11</td>
<td>4.61</td>
</tr>
<tr>
<td>Singing Abilities</td>
<td>27</td>
<td>48</td>
<td>38.21</td>
<td>5.85</td>
</tr>
<tr>
<td>Emotions</td>
<td>29</td>
<td>41</td>
<td>34.29</td>
<td>3.11</td>
</tr>
</tbody>
</table>

2.2.2. Inferential Statistics

To determine effect sizes and significances of the independent variables for predicting the JNID (see Section 2.1.1.), linear mixed-effects (LME) was fitted to the data. All statistical analyses were done on RStudio Version 1.0.153. LME was appropriate for analysing the data because it helps to protect against Type 1 errors (i.e. finding a significant effect when it is in fact not significant; Barr, Levy, Scheepers, & Tily, 2013). LME also allows for random effects, which resolve non-independence (there were multiple responses from each participant in this experiment) by assuming that there are different baselines for each participant.

Only complete staircase trials were included in the analysis, and only variables that may have some theoretical basis for predicting the JNID were included in the model.

The residuals of the LME model were homoscedastic and the Q-Q plot did not violate LME assumptions (see Appendix B).

The independent variables included in the LME model were Interval Size, Interval Direction (ascending or descending), Absolute Detune Start Value, and Staircase Direction (i.e. which direction the staircase stepped after a correct response).
A maximal random effects structure (as recommended by Barr et al. [2013]) was attempted with *Interval Size*, *Interval Direction*, *Absolute Detune Start Value*, and *Staircase Direction* in the random effects (as well as in the fixed effects), but the model failed to converge, hence only the intercept was included as a random effect. The subsequent models therefore assume that the independent variables do not vary across participants in the population, but participants’ ability overall does vary.

The variables included in the fixed effects are:

- **Interval Size**: the reference category (i.e. the category that all other categories in that variable would be compared to) being *Interval Size 1* (a minor 2\(\text{nd}/1\) semitone between two notes), as this was the smallest interval in the experiment;
- **Interval Direction**: the reference category was *Interval Direction Descending*;
- **Absolute Detune Start Value**: the reference category was 0 cents;
- **Staircase Direction**: the reference category was –10 cents;
- **Gold-MSI Active Engagement**;
- **Gold-MSI Perceptual Abilities**;
- **Gold-MSI Musical Training**;
- **Gold-MSI Singing Abilities**;
- **Gold-MSI Emotions**;
- **Absolute Pitch**: coded as 0 (does not possess absolute pitch, \(n = 22\)) or 1 (has absolute pitch, \(n = 6\)), with 0 as the reference category;
- **Age**;
- **Gender**: a categorical variable, coded as either 0 (female) or 1 (male), with 0 as the reference category.
The 5 factors of the Gold-MSI were able to be included separately because they were not multicollinear (see the Variance Inflation Factor [VIF] table in Appendix C). The Musical Training factor was of particular interest because of previous research suggesting that information stored in LTM is from previously learnt material (Berz, 1995), and as this study involves STM-LTM comparisons of intervals, the LTM of an interval would have to be previously learnt through musical training.

The LME model contained the absolute value of the JNID of the complete trials as a function of the variables listed above.

Using classical t-tests to indicate significance in mixed effects models tend to be anti-conservative/give incorrectly low p-values (Fox & Weisberg, 2015). Hence, theoretical likelihood ratio tests (LRT) were used to gain p-values for the predictors by calculating the logarithm of the likelihood ratio between a model with that particular predictor and a model without that predictor. This was done with the drop1 function in R. The results of this analysis can be seen in Table 3.
Table 3

*Variables Predicting the JNID*

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Estimate</th>
<th>Standard Error (SE)</th>
<th>$\chi^2$</th>
<th>$p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>0.53</td>
<td>0.14</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Interval Characteristics</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Interval Size 2</td>
<td>-0.07</td>
<td>0.02</td>
<td>8.57</td>
<td>.003**</td>
</tr>
<tr>
<td>Interval Size 3</td>
<td>-0.03</td>
<td>0.02</td>
<td>1.41</td>
<td>.234</td>
</tr>
<tr>
<td>Interval Size 4</td>
<td>-0.04</td>
<td>0.02</td>
<td>2.38</td>
<td>.123</td>
</tr>
<tr>
<td>Interval Size 5</td>
<td>-0.06</td>
<td>0.02</td>
<td>5.81</td>
<td>.016 *</td>
</tr>
<tr>
<td>Interval Size 6</td>
<td>-0.02</td>
<td>0.03</td>
<td>0.27</td>
<td>.605</td>
</tr>
<tr>
<td>Interval Size 7</td>
<td>-0.07</td>
<td>0.03</td>
<td>6.28</td>
<td>.012 *</td>
</tr>
<tr>
<td>Interval Size 8</td>
<td>-0.01</td>
<td>0.03</td>
<td>0.10</td>
<td>.756</td>
</tr>
<tr>
<td>Interval Size 9</td>
<td>-0.03</td>
<td>0.02</td>
<td>2.30</td>
<td>.129</td>
</tr>
<tr>
<td>Interval Size 10</td>
<td>-0.07</td>
<td>0.02</td>
<td>8.22</td>
<td>.004 **</td>
</tr>
<tr>
<td>Interval Size 11</td>
<td>-0.04</td>
<td>0.02</td>
<td>2.14</td>
<td>.143</td>
</tr>
<tr>
<td>Interval Size 12</td>
<td>-0.06</td>
<td>0.02</td>
<td>4.68</td>
<td>.031 *</td>
</tr>
<tr>
<td>Interval Direction</td>
<td>0.00</td>
<td>0.01</td>
<td>0.03</td>
<td>.865</td>
</tr>
<tr>
<td><strong>Absolute Detune Start Value</strong></td>
<td>0.04</td>
<td>0.01</td>
<td>16.96</td>
<td>&lt; .001 ***</td>
</tr>
<tr>
<td><strong>Staircase Direction</strong></td>
<td>-0.01</td>
<td>0.01</td>
<td>0.52</td>
<td>.470</td>
</tr>
<tr>
<td><strong>Musical Experience</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gold-MSI Active Engagement</td>
<td>0.00</td>
<td>0.00</td>
<td>1.40</td>
<td>.237</td>
</tr>
<tr>
<td>Gold-MSI Perceptual Abilities</td>
<td>0.00</td>
<td>0.00</td>
<td>0.76</td>
<td>.383</td>
</tr>
<tr>
<td>Gold-MSI Musical Training</td>
<td>-0.01</td>
<td>0.00</td>
<td>12.49</td>
<td>&lt; .001 ***</td>
</tr>
<tr>
<td>Gold-MSI Singing Abilities</td>
<td>0.00</td>
<td>0.00</td>
<td>0.22</td>
<td>.636</td>
</tr>
<tr>
<td>Gold-MSI Emotions</td>
<td>0.00</td>
<td>0.00</td>
<td>1.64</td>
<td>.201</td>
</tr>
<tr>
<td><strong>Demographics</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Absolute Pitch</td>
<td>-0.01</td>
<td>0.03</td>
<td>0.19</td>
<td>.666</td>
</tr>
<tr>
<td>Age</td>
<td>0.00</td>
<td>0.00</td>
<td>2.04</td>
<td>.153</td>
</tr>
<tr>
<td>Gender</td>
<td>0.03</td>
<td>0.02</td>
<td>1.80</td>
<td>.180</td>
</tr>
</tbody>
</table>

*Note. * $p < .05$, ** $p < .01$, *** $p < .001$. The estimate and SE values are in semitones.*
For Interval Size, the JNID for each of the 12 intervals separate can be seen in Figure 4. It was found that judging the intonation of a major 2\textsuperscript{nd} (Interval Size 2) rather than a minor 2\textsuperscript{nd} (Interval Size 1) made the JNID smaller by 6.57 cents, $\chi^2(1) = 8.57$, $p = .003$.

Judging the intonation of a perfect 4\textsuperscript{th} (Interval Size 5) rather than a minor 2\textsuperscript{nd} made the JNID smaller by 5.53 cents, $\chi^2(1) = 5.81$, $p = .016$.

Judging the intonation of a perfect 5\textsuperscript{th} (Interval Size 7) rather than a minor 2\textsuperscript{nd} made the JNID smaller by 6.53 cents, $\chi^2(1) = 6.28$, $p = .012$.

Judging the intonation of a minor 7\textsuperscript{th} (Interval Size 10) rather than a minor 2\textsuperscript{nd} made the JNID smaller by 6.79 cents, $\chi^2(1) = 8.22$, $p = .004$.

Judging the intonation of an octave (Interval Size 12) rather than a minor 2\textsuperscript{nd} made the JNID smaller by 6.24 cents, $\chi^2(1) = 4.68$, $p = .031$. 
Figure 4. Average of the absolute value of the JNID for each interval size, with bootstrapped 95% confidence intervals derived from 10,000 samples.

Pairwise comparisons of Interval Size were performed with Bonferroni correction on R using the anova function. As there were a total of 66 possible comparisons, the alpha level was \(0.05/66 = 0.00076\). Subsequently, no other levels of Interval Size were significantly different from each other.

The JNID was significantly worse by 3.58 cents when Absolute Detune Start Value was 50 cents compared to when it was 0 cents, \(\chi^2(1) = 16.96, p < .001\).

Gold-MSI Musical Training had a significant effect on the JNID, where for one unit increase in Gold-MSI Musical Training, the JNID would decrease by 0.81 cents, \(\chi^2(1) = 12.49, p < .001\).
2.3. Discussion

There were four key findings from this experiment:

- There was little variation in the sample with regards to the average JNID from each participant;
- The Gold-MSI factor of Musical Training significantly and inversely predicted the JNID;
- Certain intervals had a significantly lower JNID;
- An interval starting out of tune in the staircase resulted in a significantly larger JNID.

2.3.1. Small Variation of Average Just Noticeable Interval Differences and Musical Training

The small variation in average JNIDs between participants indicates consistency across participants, and may be linked with the Gold-MSI factor *Musical Training* being significant and inversely proportional to the JNID. *Musical Training* refers to how much musical training and practice on a musical instrument an individual has had (Müllensiefen et al., 2014). Since a higher *Musical Training* score resulted in a smaller JNID, and since this experiment involved STM-LTM comparisons of an interval, and previously learnt material such as musical training is stored in LTM (Berz, 1995), this suggests that a higher *Musical Training* score and a correspondingly smaller JNID may be because of more interval information stored in LTM, and better retrieval of from LTM. This concurs with past research finding that musical training is important for recognising intervals (Cuddy & Cohen, 1976).

The range of average JNIDs from each participant found in the present experiment (13.7 to 32.5 cents) is similar to other ranges found in other experiments on adjustments of melodic intervals. For example, adjustment methods have found just noticeable differences in the range of 10 to 30 cents (Burns & Ward, 1978), even though they did not use an adaptive
staircase procedure like the one that was used in the present experiment. Rather, a frequency oscillator was used to adjust the second note of the standalone interval (Rakowski, 1990). The approximately same range of JNIDs found in the adjustment/frequency oscillator experiments and in the current staircase experiment provides a strong support for the consistency of the range of JNIDs for the 12 chromatic intervals inclusive within an octave.

2.3.2. Interval Sizes Influencing the Just Noticeable Interval Difference

The interval sizes that resulted in significantly smaller JNID values can be hypothesised by some intervals being more common in western music and thus more familiar, thus potentially making them easier to judge the intonation of to result in a smaller JNID. This would be the case for the major 2\textsuperscript{nd} interval, which had the significantly smallest JNID compared to the minor 2\textsuperscript{nd}. Small intervals tend to be present more frequently in melodies than larger ones, such that for intervals larger than 5 or 6 semitones, their occurrence in melodies exponentially decreases (Vos & Troost, 1989). Furthermore, when analysing melodies composed by European composers (e.g. a range of classical composers, and the Beatles) and folk music for how frequently the 13 chromatic intervals (unison to an octave) occurred in them regardless of interval direction (ascending or descending), it was found that the major 2\textsuperscript{nd} interval was the most frequently occurring interval (Vos & Troost, 1989). This may then explain why the major 2\textsuperscript{nd} resulted in a significantly better JNID than the minor 2\textsuperscript{nd}.

However, the hypothesis that smaller intervals have a smaller JNID because they occur more frequently, does not explain why the smallest JNID was observed for the minor 7\textsuperscript{th}. This can be hypothesised though with the notion of interval class, which is a term for all the possible intervals two pitch classes can create (Forte, 1973). For example, C and D can create an interval of a major 2\textsuperscript{nd}, or a minor 7\textsuperscript{th}, or a major 9\textsuperscript{th}, etc. In relation to the current experiment, the minor 7\textsuperscript{th} is in the same interval class as the major 2\textsuperscript{nd} (it is the inverse of the
major 2\(^{nd}\), and both significantly predicted the JNID. However, this does not follow through to the minor 3\(^{rd}\) and major 6\(^{th}\) for example, which did not have similar JNID values. Hence, interval classes seem to only partially explain the data.

As for the perfect 4\(^{th}\), perfect 5\(^{th}\), and octave (the remaining intervals that significantly resulted in a better JNID than the minor 2\(^{nd}\)), neither familiarity/frequency of occurrence nor interval classes explain why these intervals had a significantly smaller JNID than the minor 2\(^{nd}\). Nor does the previous research discussed thus far explain why the minor 2\(^{nd}\) had the largest JNID when, with the frequency of occurrence explanation (smaller intervals occur more frequently than larger intervals), the minor 2\(^{nd}\) should also have a small JNID. However, JNFDs found for harmonic intervals (i.e. two notes of an interval heard/presented at the same time) may match more closely to the findings of the current experiment. For example, harmonic intervals had average deviations in tuning adjustments that ranged from 13.5 cents for a perfect 4\(^{th}\), to 22 cents for a tritone (Moran & Pratt, 1926). This can be explained with the consonance of intervals, which is where for certain consonant intervals (e.g. the perfect 4\(^{th}\), perfect 5\(^{th}\), octave) the harmonics of each note of the interval coincide with the harmonics and fundamental frequency of the other note, therefore giving the impression that the notes are fused together (DeWitt & Crowder, 1987). This makes it easier to tell if consonant intervals are in tune or not, because if they are not the frequencies of each note of the interval start to beat with each other (i.e. the periodic ‘beat’ that two sound waves make when they interfere with each other is heard), and if they are in tune the beating is not present. As consonance is for harmonic intervals, the comparable concept for melodic intervals is called tonal affinity.

Tonal affinity is the perception of how well successive tones fit together (Milne, Laney, & Sharp, 2016). A psychoacoustic factor that affects affinity is called spectral pitch similarity, which describes the similarity of tones based on their frequencies and amplitudes.
of partials (Milne et al., 2016). Intervals (not including unison) that have the highest spectral pitch similarity are the octave, perfect 4th, and perfect 5th (Milne et al., 2016), but these results may only partially explain the findings of the current experiment. This is because, with spectral pitch similarity, only the octave, perfect 4th, and perfect 5th would have lower value JNIDs, but the results of this experiment showed that the major 3rd for example also had a low JNID value.

In summary, the familiarity of intervals does not fully explain the results of this experiment because with familiarity, one would expect to see a positive slope to the data. Nor do interval classes explain the data completely, since not all intervals that were in the same interval class had similar JNID values. Likewise, spectral pitch similarity does not fully explain the results of this experiment, because one would expect to see only the octave, perfect 4th, and perfect 5th with lower JNID values compared to the other intervals. Future studies could investigate reasons as to why the results of the current experiment occurred.

2.3.3. Absolute Detune Start Value Influencing the Just Noticeable Interval Difference

The reason why an Absolute Detune Start Value of 50 cents compared to 0 cents led to a significantly higher JNID is not entirely clear. There are some superficially plausible explanations. The first is from a categorical pitch perspective, where if the starting point of a staircase is likened to a tonal melodic environment, or a tonal context (e.g. Rakowski, 1990; Warrier & Zatorre, 2002), then the first interval that is heard in the staircase would be used to judge all subsequent presented intervals against it, by using the first interval as a reference with which to categorise subsequent pitches heard. If the first interval in the staircase is out of tune (an Absolute Detune Start Value of 50 cents), then the tonal context is out of tune, and all subsequent intervals heard would be judged for intonation in reference to an out of tune interval, which may lead to greater inaccuracies in intonation with regards to 12-TET. In comparison, when the first interval heard in the staircase has an Absolute Detune Start Value
of 0 cents, it is already in tune with 12-TET, hence leading that staircase to end with a smaller JNID. However, this explanation is not sufficient because in the experiment, each step of progressing through each staircase was presented in a random order. This means that after hearing a particular interval, it would be 16 to 31 intervals later before that interval from the same staircase would be heard again. It is extremely unlikely that one could remember the previous step of a staircase with so many intervening items causing loss of pitch information in STM, which has a capacity of 7 to 11 notes (Pembrook, 1987). Therefore, the design of the experiment makes it unlikely that differently tuned tonal contexts/starting points of staircases explains the significant differences in JNIDs.

The last possible explanation is that at some point midway through the experiment, several staircases will finish, allowing the second iteration of those just-finished staircases to start. Some of these staircases will have an Absolute Detune Start Value of 0 cents, which will provide an accurate 12-TET reference point for the intervals from other staircases shortly heard thereafter, particularly those that are about to finish. However, as the staircases are progressed through once again, those staircases which began at 0 cents will be less likely to be still at 0 cents, and item decay will increasingly have more of an effect, such that the intervals that began not out of tune (0 cents) in the staircase will increasingly be forgotten. While this explanation is arguably more likely than the previous two offered, there was no significant difference found between the JNIDs from the two staircase numbers (either 1 or 2, first iteration or second iteration), $\chi^2(1) = 1.20$, $p = .27$, which therefore does not support this explanation.

This finding is therefore at the moment unexplained, and unexpected, since the experiment was designed in such a way (i.e. mixing up the steps of every staircase rather than hearing and completing one staircase at a time) in order to avoid this effect. Future
experiments could attempt to try and replicate this finding and/or investigate why this result occurred.


The aim of this experiment was to investigate to what extent pitch memory was affected by time decay when remembering and reproducing a single pitch. Two tasks were designed to address this, Task 1 being Single Pitch Reproduction, where participants could reproduce the single pitch immediately after its presentation on the slider, and Task 2 being Single Pitch Reproduction After a Delay where participants could only reproduce the single pitch 6 seconds after its presentation on the slider. The 6 seconds before pitch reproduction had no interfering notes, only silence, and it was 6 seconds long because 6 seconds plus the 1-second-long single pitch was the length of a melody in Experiment 3. Experiment 2 therefore provides a baseline for Experiment 3 in which there are no interfering notes that may interfere with reproduction of a single pitch.

3.1. Method

3.1.1. Design

Experiment 2 was a quantitative within-subjects design consisting of two tasks: Task 1 (single pitch reproduction) and Task 2 (single pitch reproduction after a delay – 6 seconds of silence). Task 1 was to gauge participants’ ability at reproducing a pitch on the slider without any time decay (silence) or interfering notes, so as to provide a baseline for their pitch reproduction/matching skills. Task 2 was to gauge the effect of time decay on participants’ ability to reproduce a pitch on the slider.

The independent variable was Decay/Silence (2 levels: none, 6 seconds), a categorical variable which indicated the only difference between the two tasks in this experiment: in
Task 1 the presented pitch could be reproduced immediately after its presentation, while in Task 2 the presented pitch could only be reproduced after 6 seconds of silence.

The dependent variables were Pitch Error (pitch reproduction accuracy, measured in cents) and Response Time (the time taken to reproduce the pitch, measured in milliseconds). Pitch Error was the absolute difference between the presented pitch and the participant’s reproduced pitch on the slider. Response Time was measured differently in the two tasks: in Task 1 it was measured from the end of the pitch presentation to the time the participant submitted their reproduced pitch on the slider; in Task 2 it was measured from the end of the 6 seconds of silence to the time the participant submitted their reproduced pitch on the slider. This would make Response Time of the two tasks comparable.

3.1.2. Participants

Participants were 21 musicians (15 females, 6 males) aged 18 to 45 years old ($M = 27.9$ years, $SD = 7.6$), with 5 to 23 years of formal training ($M = 11.8$ years, $SD = 4.7$) on a musical instrument (which included voice). Participants were recruited via social media, flyers put up on WSU campuses, WSU CareerHub (a university website for job advertisements), and from the MARCS Institute for Brain, Behaviour and Development Music Cognition and Action participant database. Several participants from Experiment 1 were also contacted again with information about Experiment 2 and 3 (which were run one after another in one session) as they had previously expressed interest in being contacted for the follow-up study. As such, 12 participants took part in Experiment 1, 2 and 3. All participants resided in Sydney and were reimbursed $20/hr to cover travel expenses.

Since the results of Experiment 1 were used to guide features of Experiment 3, and since Experiment 2 and Experiment 3 were run in one session, the same eligibility criteria from Experiment 1 were kept for Experiment 2: participants were, (a) aged 18 to 45 years
old, and (b) musicians with 5 or more years of formal training on a musical instrument (which included voice).

3.1.3. Apparatus

This experiment was run similarly to Experiment 1; participants individually used a Mac laptop with a computer mouse to progress through the experiment presented on the laptop. All auditory stimuli were presented in mono and heard binaurally through closed back over the ear headphones. The experiment was constructed and run on Max/MSP (Cycling ’74, San Francisco, CA), which is a program that can create audio, video and graphics for user interaction.

3.1.3.1. Slider. The slider was an onscreen device (see Figure 5) that was programmed to cover a range of two octaves from A4 to A6, as this encompassed the range of all the single pitches that would be presented in this experiment (see Section 3.1.4.1. for the range of single pitches). The reason why it was not the exact range of pitches was because when participants would have to reproduce the lowest and highest pitch, if the extreme ends of the slider matched these pitches, it would be quite easy to reproduce them accurately. The slider was 100 pixels in height and 1679 pixels in length, which gave it a resolution of 1 pixel to approximately 1.43 cents. This resolution allowed movement on the slider to be perceived as a continuous output of pitch. The length of 1679 pixels was the maximum length the slider could be for a computer with a screen resolution of 1680x1050 pixels. The slider was controlled by the computer mouse: clicking on the slider for the first time would cause the sound to start (the pitch of the sound would correspond to where the click was registered) and a marker (a thin line) to appear that indicated where the slider was currently positioned. The sound would continue until the user clicked the ‘Submit Response’ button just below the slider. The slider sound was the same that was used in Experiment 1 (sawtooth waves put through a lowpass 3rd order (−18dB per octave) Butterworth filter, with a cut-off frequency of
2756.25 Hz). This same sound was used for presenting all single pitch stimuli in Experiment 2.

Figure 5. A screenshot of how participants were first presented and introduced to the slider.

3.1.3.2. Musical experience questionnaire. The Gold-MSI musical experience questionnaire was hosted on Qualtrics.

3.1.4. Materials

3.1.4.1. Single pitches. Single pitches were all tuned to 12-TET, and were all the semitones from Bb4 to G6 inclusive, 22 pitches in total. This was the same range covered by the interval stimuli in Experiment 1, and the same range as all the melodies in Experiment 3. The variable name Note describes the name of each pitch. The duration of each pitch was 1 second long. All participant responses for both tasks were recorded on the computer using Max/MSP.
3.1.4.2. **Musical experience questionnaire.** The Gold-MSI (Müllensiefen et al., 2014) collected information on participants’ musical experience and expertise (see Appendix A), for the same reasons as in Experiment 1 (see Section 2.1.4.2.). The Gold-MSI also provides an overall score of musical expertise called *General Sophistication* (Müllensiefen et al., 2014).

3.1.5. **Procedure**

Task 1 and Task 2 were completed on the computer. All the instructions for the tasks were displayed onscreen. All participants completed the two tasks in the same order.

Participants were first introduced to the slider and were given as much time as they needed to familiarise themselves with using it, for example producing a pitch and seeing how the pitch changed (increased/decreased) over the length of the slider. They were then given a demonstration of how they would use the slider in the experiment. In the demonstration, the pitch F6 was sounded, followed by a demonstration of how to reproduce that pitch on the slider, with instructions to either slide (holding down the mouse button) or click (at different points on the slider) to fine-tune the reproduced pitch. Following this was a practice trial for further familiarisation with the slider, where participants heard an F5 pitch, and this time they were instructed to reproduce it by themselves.

In both Task 1 and Task 2, participants were informed that they should reproduce the pitch as accurately and as quickly as possible, and were instructed to not overthink their response. They were also instructed not to sing, hum, or articulate in any way throughout the task, as this may affect their pitch reproduction ability, and that they could only hear each pitch/melody once. A progress bar was displayed at the top of the screen so that participants could track their progress.

3.1.5.1. **Task 1: Single pitch reproduction.** Participants were instructed onscreen to “Reproduce the pitch on the slider!”; this text appeared above the slider immediately after the
presentation of each pitch. They then reproduced the pitch. The ‘Submit Response’ button only appeared after the slider was clicked/produced a sound. Participants were first presented with one practice trial of a D6 pitch. Then followed the experimental trials: 22 pitches (Bb4 to G6 inclusive) presented in a random order, one at a time in separate trials.

3.1.5.2. Task 2: Single pitch reproduction after a delay. Participants were again instructed onscreen to “Reproduce the pitch on the slider”, the text like before appearing above the slider, but this time the slider (and text) only appeared after the 6 seconds of silence. As before, participants then reproduced the pitch, and the ‘Submit Response’ button only appeared after the slider was clicked/produced a sound. Participants were first presented with one practice trial of a G5 pitch. Then followed the experimental trials: 22 pitches (Bb4 to G6 inclusive) presented in a random order, one at a time in separate trials.

To ensure accurate comprehension of the instructions, participants were asked at the end of each set of instructions for each task to repeat back in their own words what they had to do for the upcoming task. Participants were also told that they could take a break whenever they needed to, and that they could ask the experimenter any questions they had at any time.

After completing Experiment 2, participants immediately and in the same testing session moved on to Experiment 3.

3.2. Results

Experiment 2 was made up of Task 1 (single pitch reproduction) and Task 2 (single pitch reproduction after a delay). The absolute value of pitch error was used in the analyses of both these tasks. The absolute pitch error was calculated by subtracting the pitch of the note to be reproduced on the slider (in MIDI number format, which can have decimal values) from the response (also in MIDI number format), to give a value in semitones (an increase of one MIDI number is 100 cents). Absolute pitch error was used because averaging positive and
negative pitch errors would not give an accurate value for the average amount of pitch error (e.g. negative and positive values may balance each other out to give an inaccurate average).

Participants were not excluded from the analysis based on how they performed in Experiment 2. However, in Experiment 3, two participants were excluded for reasons detailed in Section 4.2, and since Experiment 2 and 3 were run one after another, those two participants were also excluded from Experiment 2. Two more participants were recruited to replace their data in both Experiment 2 and 3.

3.2.1. Descriptive Statistics

3.2.1.1. Task 1: Single pitch reproduction. The absolute pitch error for this task from 462 trials ranged from 0.06 to 441.75 cents ($M = 18.83$ cents, $SD = 29.36$ cents). Participants’ individual average absolute pitch error for this task ranged from 9.68 to 77.67 cents ($M = 18.83$ cents, $SD = 14.39$ cents).

3.2.1.2. Task 2: Single pitch reproduction after a delay. The absolute pitch error for this task from 462 trials ranged from 0.24 to 639.07 cents ($M = 28.08$ cents, $SD = 66.11$ cents). Participants’ individual average absolute pitch error for this task ranged from 7.14 to 100.63 cents ($M = 28.08$ cents, $SD = 25.48$ cents).

See Figure 6 for a graph of average absolute pitch error of Task 1 and 2.

3.2.2. Inferential Statistics

There were two main aims for the analysis of Experiment 2: to address the hypothesis that the silence of 6 seconds before being allowed to reproduce the presented pitch would increase pitch error (i.e. pitch error in Task 2 being significantly worse than in Task 1), and to investigate if there were any other significant predictors that informed pitch error based on past research; for example, pitch discrimination being better at certain frequencies (Moore, 1973; Zeitlin, 1964). Investigating if there are other predictors that significantly affect pitch error may inform which variables should be included in the analysis of Experiment 3.
As before, the statistical analyses were done on RStudio Version 1.0.153, and the graphs were drawn on Matlab R2016b.

The data was first fitted to a LME model, but the residuals were heteroscedastic and skewed, so a generalised linear mixed effects (GLME) model with gamma distribution and log link was used instead. The log link meant that if there were any significant predictors, the exponential of the estimate of a significant predictor would be the multiplying factor by which the absolute pitch error would change for each unit increase of that significant predictor, all else being equal. For example, if \( x \) was the estimate of a significant predictor, then absolute pitch error would change by \( 100^\% (\exp(x) - 1)^\% \) for every one unit increase of the significant predictor.

A maximal random effects structure was attempted (i.e. having all fixed effects included as random effects; in Task 1 and 2 this would be the variable \( Note \) because each participant had data for 22 trials per task), but the model did not converge. Therefore, only the intercept was included as a random effect, which meant that subsequent models assumed that independent variables do not vary across participants in the population, but participants’ ability overall does vary.

The fixed effects included in the model were:

- \( Decay/Silence \): the reference category was \( Decay/Silence: None \) (Task 1);
- \( Note \): the pitch as a MIDI number that was presented to the participant for them to reproduce on the slider;
- \( Gold-MSI General Sophistication \): an overall score of musical expertise.

Participant scores ranged from 81 to 123 (\( M = 101.05, SD = 11.08 \)). The 5 separate factors of the Gold-MSI were not able to be included because of multicollinearity (see VIF table in Appendix D);
• *Absolute Pitch*: coded as 0 (does not possess absolute pitch, \(n = 17\)) or 1 (has absolute pitch, \(n = 4\)), with 0 as the reference category;

• *Age*;

• *Gender*: a categorical variable, coded as either 0 (female) or 1 (male), with 0 as the reference category.

Value ranges varied widely between certain variables. Certain variables had a range of values that were much bigger in value than other variables (e.g. scores ranging from 80 to 120 vs. scores ranging from 0.1 to 1). This caused the `lme4` function in R to give warning messages. Hence, the variables with larger scores were rescaled by dividing their scores by 100 to make the values for all the variables more in the same range. To interpret their estimates, the estimate would be divided by 100. The variables that were subject to this adjustment were: *Note, Gold-MSI General Sophistication*, and *Age*.

*Note* squared was attempted to be used in the model to test for non-linearities in absolute pitch error, but the term was not significant, so it was removed.

The GLME was subsequently run, containing the absolute pitch error of the trials in Task 1 and 2 as a function of the variables listed above. See Appendix E for the residual plot of this model.

Using classical \(t\)-tests to indicate significance in mixed effects models tend to be anti-conservative/give incorrectly low \(p\)-values (Fox & Weisberg, 2015). Hence, theoretical LRTs were used to gain \(p\)-values for the predictors by calculating the logarithm of the likelihood ratio between a model with that particular predictor and a model without that predictor. This was done with the `drop1` function on R. The results of this analysis can be seen in Table 4.
Table 4

Variables Predicting the Absolute Pitch Error in the Single Pitch Reproduction and Single Pitch Reproduction After a Delay Tasks

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Estimate</th>
<th>SE</th>
<th>$\chi^2$</th>
<th>$p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>2.45</td>
<td>0.01</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Decay/Silence</td>
<td>0.25</td>
<td>0.01</td>
<td>17.78</td>
<td>&lt; .001 ***</td>
</tr>
<tr>
<td>Note</td>
<td>-3.05</td>
<td>0.01</td>
<td>42.36</td>
<td>&lt; .001 ***</td>
</tr>
<tr>
<td>Gold-MSI General Sophistication</td>
<td>-1.46</td>
<td>0.01</td>
<td>2.81</td>
<td>.094</td>
</tr>
<tr>
<td>Absolute Pitch</td>
<td>-0.36</td>
<td>0.22</td>
<td>2.19</td>
<td>.139</td>
</tr>
<tr>
<td>Age</td>
<td>-1.70</td>
<td>0.40</td>
<td>1.79</td>
<td>.181</td>
</tr>
<tr>
<td>Gender</td>
<td>-0.46</td>
<td>0.20</td>
<td>4.67</td>
<td>.031 *</td>
</tr>
</tbody>
</table>

Note. * $p < .05$, ** $p < .01$, *** $p < .001$.

Decay/Silence had a significant effect on absolute pitch error when reproducing the presented pitch, $\chi^2(1) = 17.78$, $p < .001$. Absolute pitch error significantly increased by 29% when there was 6 seconds of silence (Task 2) compared to when there was no silence (Task 1) (see Figure 6). Therefore this confirms the hypothesis of Experiment 2.
Figure 6. Average absolute pitch error in semitones for Decay/Silence: None (Task 1) and Decay/Silence: 6 seconds (Task 2), with bootstrapped 95% confidence intervals derived from 10,000 samples.

With regards to investigating if there were any other predictors that significantly affected absolute pitch error, Note had a significant effect on absolute pitch error when reproducing the presented single pitch, $\chi^2(1) = 42.36, p < .001$. Absolute pitch error significantly increased by 95% as the pitch of Note decreased. This significant effect can be seen in Figure 7.
Figure 7. Average absolute pitch error in semitones for each Note presented in Decay/Silence: None (Task 1) and Decay/Silence: 6 seconds (Task 2), with bootstrapped 95% confidence intervals derived from 10,000 samples.

Gender also unexpectedly had a significant effect on absolute pitch error when reproducing the presented single pitch, $\chi^2(1) = 4.67, p = .031$. Absolute pitch error significantly increased by 37% in females compared to males. Across Task 1 and Task 2, the range of absolute pitch error for females was 0.06 to 639.07 cents ($M = 0.27$ cents, $SD = 59.26$ cents), and for males the range was 0.18 to 243.42 cents ($M = 0.14$ cents, $SD = 17.58$ cents). The graph for this can be seen in Figure 8.
Figure 8. Absolute pitch error in semitones by Gender for Decay/Silence: None and Decay/Silence: 6 seconds combined (Task 1 and Task 2), with bootstrapped 95% confidence intervals derived from 10,000 samples.

Separating this by task, in Task 1 the range of absolute pitch error for females was 0.06 to 441.75 cents ($M = 21.56$ cents, $SD = 33.77$ cents), and the range for males was 0.18 to 56.62 cents ($M = 12.01$ cents, $SD = 9.99$ cents). In Task 2 the range of absolute pitch error for females was 0.24 to 639.07 cents ($M = 33.24$ cents, $SD = 76.32$ cents), and the range for males was 0.42 to 243.42 cents ($M = 15.18$ cents, $SD = 22.70$ cents). The graph of this can be seen in Figure 9.
Figure 9. Absolute pitch error in semitones for Gender in Decay/Silence: None (Task 1) and Decay/Silence: 6 seconds (Task 2) separately, with bootstrapped 95% confidence intervals derived from 10,000 samples.

3.3. Discussion

There were three main findings in this experiment. The first confirmed the hypothesis of this experiment: pitch error was larger when there was intervening silence (time decay) before the presented pitch could be reproduced, compared to when there was no intervening silence. The second finding was that higher frequencies of pitch resulted in smaller pitch errors. The third finding was a significant effect for gender, where overall throughout this experiment, males had on average a smaller pitch error when reproducing the pitch compared to females.
3.3.1. Time Decay Increases Pitch Error

Pitch error being greater when reproducing a pitch after a length of silence compared to when there is no silence concurs with previous research on time decay. This is where a loss of pitch information occurs in STM (which can only hold a small amount of information for a short time) when there is a delay before recall occurs (Williams, 1975). These results have also been found for verbal auditory and visual STM research (Williams, 1975). However, research that compared time decay to item decay suggests that STM for pitch in pitch discrimination tasks decays more due to item decay than time decay (e.g. Deutsch, 1970; Williams, 1975). This may be the case only for pitch discrimination tasks however, whilst pitch reproduction tasks such as the tasks in Experiment 2 may be more susceptible to time decay, potentially due to different mechanisms behind completing these tasks. Future studies could investigate a comparison of these two tasks.

3.3.2. Higher Frequencies of Pitch Decrease Pitch Error

An investigation into whether the pitch of the note being reproduced had an effect on pitch error was prompted by past research noting that a certain range of frequencies such as 490 to 990 Hz were best for pitch discrimination (Zeitlin, 1964), in combination with several participants commenting in this experiment that the higher frequencies were easier to reproduce on the slider. Subsequently, pitch error was found to decrease as the presented note increased in frequency (became higher in pitch). The range of notes used for pitch reproduction in this experiment, converted into frequencies for comparison with previous research, was 195.998 to 932.328 Hz. This covers the range that Zeitlin (1964) stated was best for pitch discrimination, and in addition to this, the higher end of the range used in the current experiment corresponds to smaller pitch error, which therefore seems to confirm Zeitlin’s range of frequencies for most accurate pitch discrimination.
The inverse proportional relationship (higher frequencies being reproduced resulting a lower pitch error) found in this experiment has also been seen in experiments on pitch discrimination, such as Moore (1973), where two tones were discriminated for whether the first or the second one was higher in pitch. The two tones were always 200 ms apart, and a range of frequencies (250, 500, 1000, 2000, 4000, 5000, 6300 and 8000 Hz) and duration of tones were tested (6.25 to 200 ms; Moore, 1973). All durations of tones showed a similar inversely proportional relationship between pitch discrimination accuracy and frequency, within the range of frequencies used in the current experiment (Moore, 1973). This suggests that a pitch discrimination task and a pitch reproduction task may share some commonalities in cognitive processing such as STM/pitch memory (e.g. Berz, 1995; Williams, 1975), since the note presented in the experiment would have to be stored in memory in order to reproduce it, and the stored pitch in memory would be compared to the pitch produced by the slider in order to determine the accuracy of reproduced note. Additionally, though not covered in the current experiment, Moore (1973) found that errors in pitch discrimination sharply increased from 4000 to 5000 Hz and onwards, which incidentally is beyond the range of most musical instruments; for example the piano has a frequency range of 27.5 to 4186 Hz. Beyond this point, frequencies are clearly difficult to discriminate.

The range of frequencies for presented pitches in this experiment thus showed a similar inversely proportional relationship to pitch error/discrimination as found in other experiments, suggesting that the frequency range used in the present experiment was reasonable for accurate discrimination and hence reproduction. Similar findings in a pitch discrimination task and a pitch reproduction task also suggest that both tasks use STM/pitch memory to differentiate pitches and to try to achieve accurate pitch reproduction respectively.

3.3.3. Effect of Gender on Pitch Error
The significant effect for gender in this experiment, where males had on average a lower pitch error throughout the experiment (both Task 1 and 2) compared to females, does not appear in Experiment 1, nor in previous research. It is probable that the effect of gender on pitch error in this experiment is coincidental, and potentially due to differences in sample size of the two genders. A more balanced future experiment may investigate this potential gender effect better.

4. Experiment 3: Pitch Reproduction of the First Note of Detuned Melodies

This experiment brought together the findings of the two previous experiments, where the melodies were detuned by an amount guided by the average JNID of 12 chromatic intervals found in Experiment 1, and Experiment 2 provided a baseline for pitch reproduction of a single note. The aim of Experiment 3 was to investigate if pitch error when reproducing the first note of a melody was larger when the detune rate was below the value of the JNID, compared to when the detune rate was on or above the value of the JNID. Participants were asked to reproduce only the first note of the melody, and to rate how confident they were of their response.

4.1. Method

4.1.1. Design

This experiment consisted of the detuned melodies task, which was a quantitative 3x3 within-subjects design. The independent variables were: Detune Size (3 levels: 12.5 cents, 25 cents, 37.5 cents) and Detune Direction (3 levels: positive, negative, baseline/no change/control). The dependent variables were Pitch Error (pitch reproduction accuracy) and Response Time (the time taken to reproduce the pitch). Pitch Error was measured in cents and was the absolute difference between the first note of the melody and the participant’s reproduced pitch of the first note. Response Time was measured in milliseconds from the end
of the melody presentation to the time the participant submitted their reproduced pitch on the slider.

The levels for *Detune Size* were determined from the average JNID of 25.30 cents from Experiment 1. This value was rounded to the nearest whole cent (25 cents) for the current experiment given that a difference of 0.30 cents would be undetectable. Additionally, the average JNID across all intervals in Experiment 1 was taken because the estimates of different intervals having significantly different effects on the JNID in Experiment 1 were small (the largest estimate was for the minor 7th, where by judging this interval’s intonation, the JNID would be 6.79 cents better than when judging a minor 2nd for its intonation), and because the melodies in this experiment were made up of various interval sizes, thus rendering it difficult and ultimately irrelevant to do a weighted average JNID for Experiment 3.

For the 3 levels of *Detune Size*, one level was chosen to be at the value of the JNID, one was chosen to be below the JNID (12.5 cents below the JNID), and one was chosen to be above the JNID (12.5 cents above the JNID). This meant that for a *Detune Size* of 12.5 cents, detuning had to start on the 2nd note of the melody; for a *Detune Size* of 25 cents, detuning had to start on the 5th note of the melody; and for a *Detune Size* of 37.5 cents, detuning had to start on the 6th note of the melody, in order to keep the total detuning consistent across all *Detune Sizes*. See Figure 10 for the different onsets of the *Detune Size* levels.
### Figure 10
The amount in cents that each note of a melody is out of tune in the different levels of onset of detuning.

<table>
<thead>
<tr>
<th>Detune Size</th>
<th>2nd Note Onset</th>
<th>5th Note Onset</th>
<th>6th Note Onset</th>
</tr>
</thead>
<tbody>
<tr>
<td>12.5 cents per note</td>
<td>0 12.5 25 37.5 50 62.5 75 Cents out of tune</td>
<td></td>
<td></td>
</tr>
<tr>
<td>25 cents per note</td>
<td>0 0 0 0 25 50 75 Cents out of tune</td>
<td></td>
<td></td>
</tr>
<tr>
<td>37.5 cents per note</td>
<td>0 0 0 0 0 37.5 75 Cents out of tune</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The values of 12.5 and 37.5 cents were also below the minimum (13.7 cents) and above the maximum (32.5 cents) average JNIDs found in Experiment 1, respectively.

Detuning was performed linearly (as seen in Figure 10), in steps/quantised by the notes of the melody. Each note in a detuned melody did not drift out of tune as it was being heard, but rather the pitch was held constant for its duration. The 3 levels of *Detune Size* and their corresponding detuning onsets resulted in all detuned melodies being out of tune by 75 cents at the end note. This does not reach the most commonly reported/average semitone (100 cents) drift of ID, but this was assumed to be acceptable as this previously reported drift generally occurred over tens of seconds (Seaton et al., 2014), whereas in the present experiment each melody lasted only 7 seconds.

*Detune Direction* by *Detune Size* meant that there were 7 categories of detuning (see Figure 11). The detuning was either positive (12.5, 25, 37.5 cents), negative (−12.5, −25, −37.5 cents), or baseline/no detuning (0 cents). Melodies with no detuning were included to
gain a baseline for participants’ ability to reproduce/match the first note of the melody after hearing 6 other interfering notes that were in tune. There were 9 melodies in each category.

<table>
<thead>
<tr>
<th>Detune Direction</th>
<th>Detune Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Negative 37.5</td>
<td>0 0 0 0 0 -37.5 -75 -37.5 cents/note</td>
</tr>
<tr>
<td>Negative 25</td>
<td>0 0 0 0 -25 -50 -75 -25 cents/note</td>
</tr>
<tr>
<td>Negative 12.5</td>
<td>0 -12.5 -25 -37.5 -50 -62.5 -75 -12.5 cents/note</td>
</tr>
<tr>
<td>Baseline 0</td>
<td>0 0 0 0 0 0 0 0 0 0 cents/note</td>
</tr>
<tr>
<td>Positive 12.5</td>
<td>0 12.5 25 37.5 50 62.5 75 12.5 cents/note</td>
</tr>
<tr>
<td>Positive 25</td>
<td>0 0 0 0 0 25 50 75 25 cents/note</td>
</tr>
<tr>
<td>Positive 37.5</td>
<td>0 0 0 0 0 0 37.5 75 37.5 cents/note</td>
</tr>
</tbody>
</table>

*Figure 11.* The 7 categories of *Detune Direction* by *Detune Size*, which show how much and in what direction each note of the melody was detuned by.
As melodies often have unique individual characteristics such as their first note/key (Note 1), No. of Unique Pitches (i.e. how many different notes were there in each melody), and Up/Down Contour (3 categories: neutral contour, positive contour, and negative contour; i.e. did the melody rise or fall or stay at the same level/neutral overall), the 9 melodies assigned to each of the 7 categories of Detune Direction by Detune Size were controlled as much as possible with regards to these melodic characteristics (see Table 5 for how the melodies were spread across these 7 categories according to their melodic characteristics). This is so that variability in the categories with regards to individual characteristics of melodies would be minimised, particularly for Note 1 since Note was found to have a significant effect on absolute pitch error in Experiment 2 (higher frequencies resulted in smaller pitch errors when reproducing a single pitch).

Participants all heard the same 63 melodies, but with different conditions assigned to each group of melodies in a cyclical fashion (see Table 6). Therefore Table 5 only shows the data for 1 participant in every 7 participants, since for every 7 participants, one cycle of melody conditions was completed. The 8th participant would receive the conditions in the same order as the 1st participant.
Table 5

*The Number of Melodies With Particular Melodic Characteristics in Each Condition*

<table>
<thead>
<tr>
<th>Condition</th>
<th>Note 1</th>
<th>No. of Unique Pitches</th>
<th>Up/Down Contour</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>C  D  Eb E  F  G  A  Bb</td>
<td>4  5  6  7</td>
<td>–1  0  1</td>
</tr>
<tr>
<td>1</td>
<td>2  0  0  1  2  2  1  1</td>
<td>0  4  4  1</td>
<td>3  3  3</td>
</tr>
<tr>
<td>2</td>
<td>2  1  0  0  2  2  1  1</td>
<td>0  4  5  0</td>
<td>3  3  3</td>
</tr>
<tr>
<td>3</td>
<td>2  1  0  0  2  2  1  1</td>
<td>0  4  5  0</td>
<td>3  3  3</td>
</tr>
<tr>
<td>4</td>
<td>2  1  0  0  2  2  1  1</td>
<td>0  5  4  0</td>
<td>3  3  3</td>
</tr>
<tr>
<td>5</td>
<td>2  0  1  0  2  2  1  1</td>
<td>1  4  4  0</td>
<td>3  3  3</td>
</tr>
<tr>
<td>6</td>
<td>2  1  0  0  2  2  1  1</td>
<td>1  4  4  0</td>
<td>3  3  3</td>
</tr>
<tr>
<td>7</td>
<td>2  0  1  0  2  2  1  1</td>
<td>0  4  4  1</td>
<td>3  3  3</td>
</tr>
</tbody>
</table>

*Note. No. of Unique Pitches* was calculated by counting every new note that occurred in each melody. Two notes an octave apart in a melody were counted as two unique pitches. The *Up/Down Contour* numbers were calculated by counting a rise between two notes in a melody as a +1, no matter the size of the interval between these two notes, and correspondingly a –1 if the interval between two notes was descending. These numbers were then added up to give the number that was either positive, negative, or zero. The positive numbers were coded as ‘1’ and the negative numbers were coded as ‘–1’.
4.1.2. Participants

The same participants were used in this experiment as in Experiment 2. Experiments 2 and 3 were run one after another in the same individual session. This session had a duration of approximately one hour. See Section 3.1.2. for information on participant recruitment and demographics.

4.1.3. Apparatus

Since this experiment followed on from Experiment 2, the apparatus was the same. The slider was used once again, as well as the same sound to present the stimuli (sawtooth waves put through a lowpass 3rd order (−18dB per octave) Butterworth filter, with a cut-off frequency of 2756.25 Hz). Max/MSP (Cycling ’74, San Francisco, CA) was again used to construct and run this experiment, and record participant responses. Qualtrics was used to record responses to the Gold-MSI musical experience questionnaire (Müllensiefen et al., 2014). See Section 3.1.3. for more details.
4.1.4. Materials

4.1.4.1. Detuned melodies. All melodies used in this experiment were sourced from a corpus of monophonic melodies of 8397 European folk songs, which had already been deconstructed into underlying musical features (such as tonality, pitch, etc.) so that the subsample that was taken from this corpus sufficiently represented the original full corpus (Herff, Olsen, & Dean, 2017). The subsample contained 110 melodies, which were mathematically and perceptually tested to confirm they were novel (not familiar) to Australian listeners (Herff et al., 2017). The use of these melodies was to ensure that any familiarity with a melody would not affect participants’ pitch memory of the first note.

Sixty-three melodies were chosen from this subsample of European folk songs according to the following selection criteria:

- Melodies had to be in a major key as this is the most common and familiar tonality in western music (Trainor & Trehub, 1994);
- Melodies could not have 2 or more consecutive repeated notes because this may affect pitch memory;
- If a melody did not have an anacrusis (pick-up to the melody), then the melody had to start on the tonic note (the tonal centre of a piece of music). If a melody had an anacrusis, then the note on the first beat of the bar/immediately after the anacrusis had to be the tonic note;
- The maximum range of any melody had to be an octave.

These adjustments were then made to the melodies:

- If there was an anacrusis to the melody, the anacrusis was removed;
- The tonic note could only be heard once, so all subsequent tonic notes in the melody were removed, leaving behind only non-tonic notes (apart from the first
note). This is so that multiplicity effects will not influence how well the starting tonic note is remembered (Ben-Haim, Eitan, & Chajut, 2014);

- The melodies were shortened to being 7 notes long, as the STM storage capacity of musical pitches is approximately 6 to 11 notes (Williams, 1975; Pembrook, 1987);

- All the notes of the melody were made into the same duration: 1 second long each. Each note’s duration and the inter-onset interval between two consecutive notes was 1 second. This meant there was no silence in between notes of the melody;

- All melodies were played at the same tempo of 60 beats per minute.

All final 7-note melodies were made sure to have their last notes on a cadence where the implied next note was the tonic, so as to avoid suggested modulations to another tonic note/key.

4.1.4.2. Confidence ratings. The confidence rating scale was used to gauge how confident participants were of their pitch reproduction of the first note of the melody in the detuned melodies task. The scale was from 1 to 4, 1 being “Not confident”, 2 being “Slightly confident”, 3 being “Fairly confident”, and 4 being “Very confident”.

4.1.4.3. Musical experience questionnaire. The Gold-MSI (Müllensiefen et al., 2014) collected information on participants’ musical experience and expertise (see Appendix A), for the same reasons as in Experiment 1 and 2 (see Section 2.1.4.2.).

4.1.5. Procedure

The experiment consisted of the detuned melodies task and the Gold-MSI musical experience questionnaire (Müllensiefen et al., 2014). Instructions for the detuned melodies task were displayed onscreen. All participants completed the detuned melodies task and the Gold-MSI questionnaire in the same order.
Participants had already been introduced to the slider in Experiment 2 in the same testing session. Therefore, participants did not need extra time to familiarise themselves with the slider again.

In the detuned melodies tasks, participants were instructed onscreen to “Reproduce the FIRST NOTE of the melody!” This text appeared above the slider, and the slider and text instruction only appeared after the presentation of each melody was completed. After the ‘Submit Response’ button was clicked, participants were asked to rate how confident they were of their response. Participants were informed that the detuned melodies task may get quite difficult, and to not worry too much if they did not remember the first note of the melody by the time they had to reproduce it. They were once again informed that they should reproduce the pitch as accurately and as quickly as possible, and were told to not overthink their response. They were told that they were not allowed to sing, hum, or articulate in any way throughout the task, as this may affect their pitch reproduction/matching ability, and that they could only hear each pitch/melody once. Participants were allowed to take a break whenever they needed to, and they were told that they could ask the experimenter any questions they had at any time. A progress bar was given at the top of the screen for participants to track their progress. Participants were not informed that some melodies may be detuned, nor were they told how much they were detuned.

Two practice trials were given first so that the participant could practice reproducing/matching only the first note of each melody that was presented. Neither of these practice melodies were detuned, nor were they presented again in the experimental trials. Participants then progressed to the experimental trials (126 melodies in total), which contained two blocks of 63 melodies (54 with detuning, 9 with no detuning) presented in a random order. Melodies and their Detune Direction by Detune Size category were identical in each block, but the ordering was randomised again between blocks. In between blocks was a
break screen with instructions for participants to feel free to take a break. When the second block began, it was ensured that the melody presented in the 64th trial was never the same as the melody presented in the 63rd trial.

As in Experiment 2, to ensure accurate comprehension of the instructions, participants were asked at the end of the set of instructions to repeat back in their own words what they had to do for the upcoming task. Participants were also told that they could take a break whenever they needed to, and that they could ask the experimenter any questions they had at any time.

After the detuned melodies task, participants completed the Gold-MSI questionnaire (Müllensiefen et al., 2014).

4.2. Results

The absolute value of pitch error was used in the analysis of Experiment 3, for the same reasons as in Experiment 2 (see Section 3.2.). The absolute pitch error units were in semitones in order to keep it in the same scale as other variables such as Note 1, which was in MIDI numbers, where one integer change of a MIDI number is one semitone.

Absolute pitch errors that were rated the maximum confidence rating per participant were used to detect outliers and/or any participant data that indicated a lack of understanding of the task and/or a non-serious attempt, since the responses rated with the maximum confidence rating had the smallest average absolute pitch errors across all participants. For all but one participant, the maximum confidence rating was ‘4’. For the one participant who did not rate any of their responses as ‘4’, their maximum confidence rating was ‘3’, and their responses that were rated as ‘3’ were used for detecting outliers and later in the analysis. One participant was excluded at this point from the data of Experiment 2 and 3 (since Experiment 2 Task 2 (single pitch reproduction after a delay) is a baseline check for Experiment 3). This participant did not appear to understand Experiment 3/appeared to rate the confidence of their
response in the reverse direction (rating ‘1’ for most confident when in fact they should have rated ‘4’). Furthermore, their absolute pitch error average from the trials in Experiment 3 that they rated a confidence ‘4’ \((n = 8)\) was 84.3 cents, and this would be considered an outlier as it was 6.39 standard deviations above the mean (see Section 4.2.1.). Another participant was excluded from the data of Experiment 2 and 3 because they only had 3 years of formal training on a musical instrument. Two more participants were recruited to replace their data, thus totalling 21 participants.

4.2.1. Descriptive Statistics

Only the trials/responses rated with the maximum confidence rating were included in the analysis in order to be reasonably sure that they were confident that they had remembered the first note of the melody, and any errors in pitch reproduction afterwards would be because of Detune Size or Detune Direction, not because they had forgotten the first note and tried to reproduce a completely different pitch.

The final number of trials/responses from the 21 participants ranged from 2 to 121 per participant \((M = 47.10, SD = 39.51)\). The average absolute pitch error of each participant for this task, from only the trials rated with their maximum confidence rating, ranged from 8.52 to 48.06 cents \((M = 27.92 \text{ cents}, SD = 8.83 \text{ cents})\).

Table 7 shows for each Detune Direction by Detune Size category the number of complete trials, the range of absolute pitch errors (in cents), the mean of absolute pitch errors, and standard deviation of absolute pitch errors across all participants.
Table 7

Descriptive Statistics of the 7 Categories of Detune Direction By Detune Size

<table>
<thead>
<tr>
<th>Detune Direction</th>
<th>Detune Size (cents)</th>
<th>n</th>
<th>Minimum (cents)</th>
<th>Maximum (cents)</th>
<th>M  (cents)</th>
<th>SD  (cents)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Negative</td>
<td>37.5</td>
<td>133</td>
<td>0.36</td>
<td>164.86</td>
<td>25.67</td>
<td>23.71</td>
</tr>
<tr>
<td>Negative</td>
<td>25</td>
<td>139</td>
<td>0.54</td>
<td>150.98</td>
<td>27.07</td>
<td>23.16</td>
</tr>
<tr>
<td>Negative</td>
<td>12.5</td>
<td>128</td>
<td>0.54</td>
<td>260.75</td>
<td>34.33</td>
<td>31.67</td>
</tr>
<tr>
<td>Baseline</td>
<td>0</td>
<td>175</td>
<td>0.54</td>
<td>60.63</td>
<td>16.36</td>
<td>12.71</td>
</tr>
<tr>
<td>Positive</td>
<td>12.5</td>
<td>128</td>
<td>1.97</td>
<td>233.71</td>
<td>37.92</td>
<td>27.39</td>
</tr>
<tr>
<td>Positive</td>
<td>25</td>
<td>148</td>
<td>0.54</td>
<td>160.93</td>
<td>28.70</td>
<td>25.50</td>
</tr>
<tr>
<td>Positive</td>
<td>37.5</td>
<td>138</td>
<td>0.60</td>
<td>192.14</td>
<td>28.78</td>
<td>30.09</td>
</tr>
</tbody>
</table>

4.2.2. Inferential Statistics

The first hypothesis was that—due to RIM—if a melody is detuned below the value of the JNID, the absolute pitch error when reproducing the first note of the melody would be larger compared to when the melody is detuned at the value of the JNID, and absolute pitch error for when the melody is detuned at the value of the JNID would be larger than the absolute pitch error for when the melody is detuned above the value of the JNID.

The second hypothesis was—due to RIM and PIM—that absolute pitch error when reproducing the first note of the melody would be larger in the presence of any size of detuning, compared to when there was no detuning.

In analysing Experiment 3, only the responses/trials that had been rated with the participant’s maximum confidence rating were included ($n = 989$).

Once again the statistical analyses were done on RStudio Version 1.0.153, and the graphs were drawn on Matlab R2016b.
The data was first attempted to be fitted to a LME, but like in the analysis for Experiment 2, the residuals were heteroscedastic and skewed. Hence, a GLME with gamma distribution and log link was used instead. Interpreting the GLME’s estimates was the same as described in Section 3.2.2..

A maximal random effects structure was attempted where the five variables (Detune Direction, Detune Size, Up/Down Contour, No. of Unique Pitches, Note 1) with multiple responses per participant were included as random effects, but the model failed to converge. Therefore only the intercept was included as a random effect, which meant that subsequent models assumed that independent variables do not vary across participants in the population, but participants’ ability overall does vary.

The fixed effects included in the model were chosen because they had some theoretical basis for how/why they may contribute to absolute pitch error when reproducing the first note of a melody:

- **Detune Direction**: the reference category was no detuning/baseline;
- **Detune Size**: the reference category was 0 cents;
- **Decay/Silence 6 Seconds Average**: the average absolute pitch error for Experiment 2 Task 2 (single pitch reproduction after a delay) that each participant received. This variable was included to see if it had any effect on the absolute pitch error in Experiment 3, since it was the baseline to the baseline category of Detune Direction by Detune Size;
- **Up/Down Contour**: neutral contour was the reference category. This variable was included because Note had a significant effect on absolute pitch error in Experiment 2, such that an upwards contour of melody may affect the absolute pitch error of reproducing the first note of the melody in Experiment 3;
- **Note 1**: included to explore if the pitch of the first note affects absolute pitch error;
- No. of Unique Pitches: included to see if more different notes in a melody would affect pitch reproduction of the first note;
- Gold-MSI General Sophistication: this overall score of musical expertise was again used in the model and not the separate 5 factors of the Gold-MSI because the 5 factors had very high VIF values (see Appendix F);
- Absolute Pitch;
- Age;
- Gender.

Interactions in the fixed effects amongst meaningful variables based on Experiment 2 were attempted, such as Detune Size x Up/Down Contour and Note 1 x Up/Down Contour, because Note was significant in Experiment 2. However, the model failed to converge again, so interactions were removed one by one with the model being refitted each time until it converged. This only happened when there were no interactions between any of the fixed effects.

JNID (each participant’s average absolute error at detecting out of tune intervals in Experiment 1) was not included in the model despite guiding the amount of detuning in each of the levels of Detune Size because: the model failed to converge with it; it was not significant, \( t(601) = 0.85, p = .397 \); and not all participants in Experiment 3 had completed Experiment 1. These missing values produced issues when attempting to run theoretical LRTs with the drop1 function in R.

As in Experiment 2, value ranges varied widely between certain variables. This caused lme4 function in R to give warning messages. The variables with the larger scores were thus rescaled by dividing their scores by 100 to make the values for all the variables more in the same range. The variables subject to this adjustment were: Note 1, Gold-MSI General Sophistication, and Age.
The GLME model contained the absolute pitch error as a function of the variables listed above. See Appendix G for the residual plot of this model.

As previously mentioned in Section 3.2.2., classical \( t \)-tests indicating significance in mixed effects models tend to be anti-conservative (Fox & Weisberg, 2015). Hence, theoretical LRTs were used to gain \( p \)-values for the predictors by calculating the logarithm of the likelihood ratio between a model with that particular predictor and a model without that predictor. This was done with the drop1 function on R. The results of this analysis can be seen in Table 8.
Table 8

Variables Predicting the Absolute Pitch Error of Reproducing the First Note of the Melody

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Estimate</th>
<th>SE</th>
<th>$\chi^2$</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>-0.03</td>
<td>0.98</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Detune Size 12.5 cents</td>
<td>0.80</td>
<td>0.08</td>
<td>82.12</td>
<td>&lt; .001 ***</td>
</tr>
<tr>
<td>Detune Size 25 cents</td>
<td>0.52</td>
<td>0.08</td>
<td>38.26</td>
<td>&lt; .001 ***</td>
</tr>
<tr>
<td>Detune Size 37.5 cents</td>
<td>0.49</td>
<td>0.08</td>
<td>33.13</td>
<td>&lt; .001 ***</td>
</tr>
<tr>
<td>Decay/Silence 6 Seconds Average</td>
<td>0.23</td>
<td>0.19</td>
<td>1.46</td>
<td>.227</td>
</tr>
<tr>
<td>Melodic Characteristics</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Up/Down Contour Positive</td>
<td>-0.11</td>
<td>0.07</td>
<td>2.16</td>
<td>.142</td>
</tr>
<tr>
<td>Up/Down Contour Negative</td>
<td>0.12</td>
<td>0.07</td>
<td>2.68</td>
<td>.102</td>
</tr>
<tr>
<td>Note 1</td>
<td>-1.53</td>
<td>1.04</td>
<td>2.17</td>
<td>.140</td>
</tr>
<tr>
<td>No. of Unique Pitches</td>
<td>0.07</td>
<td>0.04</td>
<td>2.37</td>
<td>.124</td>
</tr>
<tr>
<td>Musical Experience</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gold-MSI General Sophistication</td>
<td>-0.95</td>
<td>0.48</td>
<td>3.37</td>
<td>.066</td>
</tr>
<tr>
<td>Demographics</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Absolute Pitch</td>
<td>-0.05</td>
<td>0.11</td>
<td>0.23</td>
<td>.630</td>
</tr>
<tr>
<td>Age</td>
<td>-0.57</td>
<td>0.53</td>
<td>1.05</td>
<td>.307</td>
</tr>
<tr>
<td>Gender</td>
<td>0.15</td>
<td>0.09</td>
<td>2.83</td>
<td>.092</td>
</tr>
</tbody>
</table>

Note. * p < .05, ** p < .01, *** p < .001.

All conditions were significantly different from the baseline Condition 4.

Absolute pitch error significantly increased by 122% in Condition 3+5 compared to Condition 4, $\chi^2(1) = 82.12, p < .001$.

Absolute pitch error significantly increased by 69% in Condition 2+6 compared to Condition 4, $\chi^2(1) = 38.26, p < .001$.

Absolute pitch error significantly increased by 63% in Condition 1+7 compared to Condition 4, $\chi^2(1) = 33.13, p < .001$. 
The average absolute pitch error with error bars of the categories of *Detune Size* can be seen in Figure 12.

*Figure 12.* Absolute pitch error in semitones when reproducing the first note of the melody in each *Detune Size* category, with bootstrapped 95% confidence intervals derived from 10,000 samples.

To investigate if the different categories of *Detune Size* were significantly different from each other as per the second hypothesis, planned comparisons/pairwise tests were conducted using the `anova` function in R. This would be a total of 3 planned pairwise comparisons, hence the alpha level was \( .05/3 = .016 \).
Detune Size 12.5 cents was significantly different from Detune Size 25 cents, $\chi^2(1) = 13.83$, $p < .001$. From the estimates of the fixed effects in Table 8, the absolute pitch error was significantly larger by 32% in Detune Size 12.5 cents compared to Detune Size 25 cents.

Detune Size 12.5 cents was also significantly different from Detune Size 37.5 cents, $\chi^2(1) = 17.07$, $p < .001$, where Detune Size 12.5 cents had a 36% larger absolute pitch error than Detune Size 37.5 cents.

Detune Size 25 cents and Detune Size 37.5 cents were not significantly different, $\chi^2(1) = 0.21$, $p = .645$.

To investigate if Detune Direction had a significant effect on the absolute pitch error, a GLME similar to the first GLME was run, but this time:

- Detune Size 0 cents was removed since there was no detuning in this baseline category and so was not relevant to this model, since this model would only investigate if Detune Direction had an effect on absolute pitch error;
- Detune Size 12.5 cents was now the reference category because it had the smallest amount of detuning;
- Detune Direction was added to the model, and Detune Direction Positive was the reference category.

All other variables in this model remained the same as they were in the previous model.

Once again, classical t-tests for mixed effects models tend to be anti-conservative (Fox & Weisberg, 2015), so theoretical LRTs were used instead to gain $p$-values for the predictors by calculating the logarithm of the likelihood ratio between a model with that particular predictor and a model without that predictor. This was done with the drop1 function on R. The results of this analysis can be seen in Table 9.
### Table 9

**GLME With Detune Direction to Test If It Significantly Predicted Absolute Pitch Error**

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Estimate</th>
<th>SE</th>
<th>$\chi^2$</th>
<th>$p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>(None)</td>
<td>0.93</td>
<td>1.02</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Detune Direction</td>
<td>-0.09</td>
<td>0.06</td>
<td>2.21</td>
<td>.137</td>
</tr>
<tr>
<td>Detune Size 25 cents</td>
<td>-0.26</td>
<td>0.07</td>
<td>12.28</td>
<td>&lt; .001 ***</td>
</tr>
<tr>
<td>Detune Size 37.5 cents</td>
<td>-0.31</td>
<td>0.07</td>
<td>17.13</td>
<td>&lt; .001 ***</td>
</tr>
<tr>
<td>Decay/Silence 6 Seconds Average</td>
<td>0.13</td>
<td>0.17</td>
<td>0.61</td>
<td>.435</td>
</tr>
<tr>
<td><strong>Melodic Characteristics</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Up/Down Contour Positive</td>
<td>-0.19</td>
<td>0.08</td>
<td>5.23</td>
<td>.022 *</td>
</tr>
<tr>
<td>Up/Down Contour Negative</td>
<td>0.18</td>
<td>0.08</td>
<td>5.49</td>
<td>.019 *</td>
</tr>
<tr>
<td>Note 1</td>
<td>-1.15</td>
<td>1.17</td>
<td>0.97</td>
<td>.324</td>
</tr>
<tr>
<td>No. of Unique Pitches</td>
<td>0.06</td>
<td>0.05</td>
<td>1.26</td>
<td>.261</td>
</tr>
<tr>
<td><strong>Musical Experience</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gold-MSI General Sophistication</td>
<td>-1.17</td>
<td>0.43</td>
<td>6.16</td>
<td>.013 *</td>
</tr>
<tr>
<td><strong>Demographics</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Absolute Pitch</td>
<td>-0.07</td>
<td>0.09</td>
<td>0.64</td>
<td>.423</td>
</tr>
<tr>
<td>Age</td>
<td>-0.62</td>
<td>0.44</td>
<td>1.76</td>
<td>.184</td>
</tr>
<tr>
<td>Gender</td>
<td>0.15</td>
<td>0.07</td>
<td>3.69</td>
<td>.055</td>
</tr>
</tbody>
</table>

*Note. * $p < .05$, ** $p < .01$, *** $p < .001$. 

*Detune Direction* was not significant, $\chi^2(1) = 2.21, p = .137$.

All other variables that were indicated as significant by the above theoretical LRT were disregarded, because this model had *Detune Size 0 cents* removed in order to solely investigate if there was any effect of *Detune Direction* on absolute pitch error. Therefore, this model was only valid for investigating *Detune Direction*.

To investigate if absolute pitch error in *Decay/Silence 6 Seconds Average* was significantly different from absolute pitch error in *Detune Size 0 cents*, a Mann-Whitney *U*
test was performed since it does not assume that the data is normally distributed. The data not being normally distributed had already been established prior to running a GLME in Experiment 2 and 3. The Mann-Whitney U test detects whether the median value of one group of data is significantly different to the median of another group of data. It was found that the distribution of the two groups did not significantly differ, \( Z = -1.63, p = .104 \).

4.3. Discussion

The first hypothesis of this experiment was partially confirmed. Pitch error was significantly larger when the detune size was below the value of the JNID found in Experiment 1, compared to when the detune size was at or above the value of the JNID. However, pitch error was not significantly larger for the detune size that was at the value of the JNID compared to the detune size that was above the value of the JNID.

The second hypothesis was confirmed. Pitch error was significantly smaller for the baseline size of detuning (0 cents detuning) compared to any of the non-zero sizes of detuning.

4.3.1. Pitch Error is Significantly Affected When Detune Size is Below the Just Noticeable Interval Difference

Amongst the non-zero detune sizes, pitch error when reproducing the first note (the tonic) of the melody was larger when the detune size was below the value of the JNID, compared to when the detune size was at or above the value of the JNID. Meanwhile, for the detune sizes that were at and above the value of the JNID, their pitch errors were not significantly different. It was expected that if pitch error increased as detune size decreased, it would suggest that RIM contributes to pitch error in addition to PIM; and if there were no significant differences in pitch error amongst the non-zero detune sizes, then RIM would not contribute to pitch error in addition to PIM. Instead, the hypothesis was partially confirmed.
The pitch errors from the non-zero detune sizes can overall be explained with RIM having an effect on the JNID and pitch error of reproducing the first note, in addition to PIM.

In the detune size that was below the value of the JNID, the tuning errors of the intervals would not have been perceivable. This would suggest that corrective action (e.g. attempting to ignore the detectably out-of-tune notes, or some other method) would not have occurred. The out-of-tune intervals in the melody would have been assumed to be in tune, and this may have led to pitch error of reproducing the first note of the melody being larger than the pitch errors of the other detune sizes that were of a detectable amount. The pitch error for the detune size below the value of the JNID was additionally the largest pitch error out of all the sizes of detuning (including no detuning). This finding may be an explanation for why the drift in ID can be as large as a semitone (Seaton et al., 2014).

Similar findings with regards to large pitch errors with detune sizes that are below the threshold of perception/JNID have been found in Shepard and Jordan’s (1984) research, which was a JNFD pitch discrimination study. In this study, a detuned scale was presented, where every interval of the C major scale was widened by 8.3 cents (i.e. below the value of the JNID found in Experiment 1) so that the final note was one semitone higher than it should be (Shepard & Jordan, 1984). This was followed by the presentation of a note that was one semitone higher than the original first note (Shepard & Jordan, 1984). It was found that 63% of participants stated that the final note was identical to the first note (Shepard & Jordan, 1984). This potentially hints that STM-STM comparisons and STM-LTM comparisons may be similarly affected by sizes of detuning that are below the threshold of perception, where pitch discrimination is poorer if the detune is not perceivable.

In a detune size that was at or above the value of the JNID however, the pitch errors were not significantly different. This may have been because the tuning errors of the intervals were similarly perceivable at both detune sizes of 25 cents and 37.5 cents, and would suggest
that corrective action may have been taken to try and compensate for these perceptible tuning errors. As such, this may explain why the pitch errors of these two detune sizes were not significantly different.

In summary, RIM contributes to pitch error in addition to PIM, in all sizes of detuning except zero. It seems that perceivable tuning errors reduce pitch error when reproducing the first note of the melody, compared to tuning errors that are not perceivable, which result in the largest pitch error of all detuning sizes. It is possible that tuning errors that are even more below the value of the JNID may lead to even larger pitch errors than what was seen in the current experiment, and it is also possible that once tuning errors are perceivable, pitch error plateaus no matter how much the detune size is above the JNID value. Future studies would have to confirm this by testing a wider range of detune sizes (e.g. 5 cents, 60 cents, etc.).

4.3.2. Pitch Error is Larger When Detuning Occurs

Pitch error when reproducing the first note of a melody being larger for any size of detuning compared to no detuning can be explained by RIM contributing to pitch error in addition to PIM. When all of the intervals in the melody are in tune, pitch error when reproducing the first note of the melody was the smallest. This suggests that recent in-tune intervals facilitate pitch reproduction of the first note and leads to a smaller pitch error. This can be explained by pre-existing organisational frameworks in LTM, such as 12-TET tonality, which describes the various relationships that pitches have to one another (Berz, 1995). Tonality is arguably the most important way of organising music and it suggests that STM pitch information is processed in LTM (Berz, 1995). Hence, when a melody does not contain any detuning, it is entirely in 12-TET/in tune, and so would fit more easily into the pre-existing framework of tonality. This may then lead to a smaller pitch error when reproducing the first note of an entirely in-tune melody.
However, when the most recent intervals in the melody are not in tune, it appears that once again RIM contributes to pitch error in addition to PIM, this time with the out-of-tune intervals hindering pitch reproduction of the first note and leading to a larger pitch error compared to when there was no detuning. This can again be explained by tonality structures in LTM (Berz, 1995), where if any amount of detuning occurs, the melody no longer easily fits into the pre-existing 12-TET framework. It may then be more difficult to make accurate STM-LTM comparisons. As such, all non-zero sizes of detuning led to a larger pitch error than the baseline size of detuning (no detuning). Overall this suggests that RIM was contributing to pitch error in addition to PIM in all sizes of detuning, where in-tune intervals facilitate pitch reproduction of the first note, and out-of-tune intervals hinder pitch reproduction of the first note.

5. General Discussion

The overall aim of this research was to investigate if RIM in addition to PIM contributes to pitch error when reproducing the first note of a melody, with regards to the occurrence of ID. Three experiments explored this, starting from the investigation of the factors that may affect the JNID in Experiment 1, which was relevant to both RIM and PIM; followed by Experiment 2 which investigated what affects STM for a single pitch; and finally Experiment 3 which combined what was learnt in Experiment 1 and 2 in order to investigate how RIM contributes to pitch error in addition to PIM in different sizes of detuning that simulated the occurrence of ID. The findings collectively suggest that RIM contributes to pitch error in addition to PIM in an ID situation, and, crucially, going out of tune below the threshold of detecting tuning errors of intervals leads to the largest pitch error of all the sizes of detuning in Experiment 3. This may be analogous to the gradual drift of ID, and may explain why the drift in ID can be as much as a semitone (Seaton et al., 2014).
In Experiment 1, little variation was found between participants in terms of their average JNIDs, and the Gold-MSI factor of Musical Training was found to be the largest predictor of the JNID. These findings together suggest that interval tunings are stored in LTM and utilised in STM-LTM comparisons. The JNID also suggests that small tuning errors of intervals can accumulate over time to result in ID. Additionally, though less significantly, certain interval sizes were found to significantly affect the JNID. These are hypothesised to be because of how frequently heard the interval is in music, what interval class the interval is in, and how much affinity the two notes of an interval have with each other. None of these hypotheses fully describe the data however.

In Experiment 2, time decay was found to interfere with pitch reproduction of a single note, as did lower frequencies compared to higher frequencies. The latter finding was consistent with previous findings of higher frequencies compared to lower frequencies leading to smaller pitch errors in pitch discrimination within the frequency range used across all three experiments in this thesis (e.g. Moore, 1973). The former finding was consistent with some research on STM for pitch: some studies found that time decay affects STM for pitch (e.g. Williams, 1975), while other studies found that item decay affects STM for pitch more than time decay (e.g. Deutsch, 1970). Differences in findings may be attributed to pitch discrimination tasks being used to investigate STM for pitch in previous research (e.g. Williams, 1975; Deutsch, 1970), as opposed to the pitch reproduction task used in this thesis’ Experiment 2.

RIM was found to contribute to pitch error in addition to PIM across all detune sizes in Experiment 3, including the baseline of no detuning. In the case of no detuning, reproduction of the first note of a melody is improved because the last/all notes of the melody are in tune. When the last notes of the melody were detuned by an amount that was below the threshold of perception (the JNID value found in Experiment 1), pitch reproduction of the
first note had the largest pitch error. When the last notes of the melody were detuned by an amount that was at or above the value of the threshold of perception, pitch error in reproducing first note was not as large as when the last notes were detuned by an amount that was below the threshold of perception, but also not as small as when all the notes of the melody were in tune. Lastly, there was no significant difference in pitch reproduction of the first note of a melody when the detuning of a melody was at or above the threshold of perception. This may be because both of these detune sizes were similarly perceivable, and so corrective action could be taken in both cases in an attempt to compensate for the tuning errors of the intervals. In contrast, when the detuning of a melody was below the threshold of perception, corrective action could not be taken because the detuning was not perceived, leading to the largest pitch errors in reproducing the first note of the melody across all detuning sizes in Experiment 3.

The results of Experiment 3 appear to confirm that RIM has an effect on the JNID and pitch error of reproducing the first note of a melody, in addition to PIM. There is also an added condition: a detuning below the threshold of perception affects pitch error more than a detuning that is at or above the threshold of perception, as in the latter scenario, pitch error appears to plateau once the detuning of the melody/intervals is perceivable.

5.1. Limitations and Future Studies

As briefly aforementioned in Section 4.3.1., Experiment 3 only tested three detune sizes (12.5, 25 and 37.5 cents), so it is not known if the plateau effect of pitch error once tuning errors of intervals are perceivable will continue for larger sizes of detuning. Equally, it is not certain if pitch errors will increase as the detune size approaches 0 cents in an asymptotic/exponential fashion, or some other manner (e.g. a peak in errors and a gradual decrease back to the smallest pitch error from no detuning), since only one below-JNID
detune value was tested (12.5 cents). Future studies could investigate this issue further with different detune sizes (e.g. 5 cents, 60 cents of detuning per note, and so on).

Another limitation is that PIM by itself was not investigated, and hence its effect on the pitch error of reproducing the first note of a melody is not known. PIM may be investigated by detuning, for example, a 5-note melody in two ways: with pitch deviations of 0, +25, 0, +25, 0 cents out of tune, compared to pitch deviations of 0, +25, +50, +75, +100 cents out of tune. In both cases, every interval between successive notes is out of tune by 25 cents, but in the first case the melody does not end out of tune, whilst in the second case it does. No effect of PIM would be expected in the first case, because the pitch deviations cancel each other out, whilst in the second case an effect of PIM would be expected since the pitch deviations do not cancel each other out. Comparing these two should give an estimate of the size of the effect of PIM on its own.

Another limitation of this study is that other mechanisms that may lead to ID occurring were not investigated, such as mechanisms that account for interfering notes (the middle items of a series of notes). Future studies could explore the effect the middle items of a series may have on ID.

Another limitation of the study is that a JNID for microtonal intervals (i.e. both notes of the interval not being in 12-TET, as opposed to the first one being in 12-TET as it was in Experiment 1) was not investigated. This may be relevant to Experiment 3, where every non-zero size of detuning had by the end of the melody at least one pair of notes that was a microtonal interval. Future studies could explore the JNID for microtonal intervals (which would require them to be previously learnt and stored in LTM), compare it to the JNID found in this thesis’ Experiment 1, and possibly apply it to detuned melodies.

Other future studies could explore the use of the staircase further to determine other threshold points of the JNID, as only the 50% threshold was determined in this thesis (due to
the 1 down 1 up method). Determining other threshold points may provide evidence for categorical pitch perception, since if more threshold points are measured (e.g. the 75% threshold), then the shape of the psychometric function can be better estimated. For example, a sharp boundary being perceived when there were in fact gradual pitch changes would indicate categorical pitch perception.

6. Conclusion

Achieving accurate intonation is one of the most important goals in music performance and rehearsal, particularly in unaccompanied choirs (Ganschow, 2013). However, ID occurs often in these particular groups despite the prioritisation of intonation (Ganschow, 2013). Therefore, the aim of this project was to investigate two mechanisms that may account for ID: RIM and PIM. RIM described a method of judging intonation by comparing the two most recent notes heard as an interval in STM to an interval in LTM. PIM described an alternate method of judging intonation, that of comparing the first note heard and the most recent note heard as an interval and comparing it to an interval in LTM. PIM was assumed to be present throughout this process of judging intonation since unaccompanied choirs are normally given a first note/reference pitch to use in order to stay in tune. Hence, this thesis investigated a novel concept: the additional contribution of RIM to ID, by detuning melodies to different degrees and reproducing its first note.

Overall, it was found that RIM contributed to pitch error in addition to PIM in all the melodies presented, regardless of the amount of detuning, and even if there was no detuning. When there was no detuning, the results suggested that pitch reproduction of the first note was facilitated by the other notes of the melody being in tune. When there was detuning in a melody, and if the detuning was below the threshold of detecting a tuning error in an interval, pitch error was largest out of all the detune sizes tested since the detuning was not
perceivable, so corrective action could not be taken. If the detuning was perceivable however, pitch error in reproducing the first note of the melody was still present, but to a lesser extent than when the detuning was below the threshold of perception, because detecting tuning errors in intervals may allow for corrective action to be taken. That is, the out of tune intervals in RIM will be attempted to be ignored, or the pitch deviation will be tracked, to aid more accurate reproduction of the first note of the melody.

This research has implications for choir rehearsal strategies as to why ID occurs. It also has implications for vocal pedagogy and singing contests in entertainment (Cao, Li, Liu, & Yan, 2008). It also contributes to arguably the start of investigating ID from the perspective of a melody being a series, by detuning notes of a melody by varying amounts in order to compare similar mechanisms to RIM and PIM.
References


Appendix A

Information on the Gold-MSI

Table A1

A Description of Each of the Factors of the Gold-MSI

<table>
<thead>
<tr>
<th>Gold-MSI Factor</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Active Engagement</td>
<td>Active musical engagement and musical activities</td>
</tr>
<tr>
<td>Perceptual Abilities</td>
<td>Cognitive musical ability, music listening skills</td>
</tr>
<tr>
<td>Musical Training</td>
<td>Extent of musical training and practice</td>
</tr>
<tr>
<td>Singing Abilities</td>
<td>Singing skills and activities</td>
</tr>
<tr>
<td>Emotions</td>
<td>Emotional responses to music</td>
</tr>
<tr>
<td>General Sophistication</td>
<td>Overall score of musical expertise</td>
</tr>
</tbody>
</table>

Table A2

Descriptive Statistics of the 5 Factors of the Gold-MSI From a Sample of 147,633 Participants

<table>
<thead>
<tr>
<th>Gold-MSI Factor</th>
<th>Scale Minimum</th>
<th>Scale Maximum</th>
<th>M</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Active Engagement</td>
<td>9</td>
<td>63</td>
<td>41.52</td>
<td>10.36</td>
</tr>
<tr>
<td>Perceptual Abilities</td>
<td>9</td>
<td>63</td>
<td>50.20</td>
<td>6.45</td>
</tr>
<tr>
<td>Musical Training</td>
<td>7</td>
<td>49</td>
<td>26.52</td>
<td>11.44</td>
</tr>
<tr>
<td>Singing Abilities</td>
<td>7</td>
<td>49</td>
<td>31.67</td>
<td>8.72</td>
</tr>
<tr>
<td>Emotions</td>
<td>6</td>
<td>42</td>
<td>34.66</td>
<td>5.04</td>
</tr>
<tr>
<td>General Sophistication</td>
<td>18</td>
<td>126</td>
<td>81.58</td>
<td>20.62</td>
</tr>
</tbody>
</table>
### Table A3

The Items in Each Subscale of the Gold-MSI

<table>
<thead>
<tr>
<th>No. in paper survey</th>
<th>Question</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td><strong>Factor 1 - Active Engagement</strong></td>
</tr>
<tr>
<td>1</td>
<td>I spend a lot of my free time doing music-related activities.</td>
</tr>
<tr>
<td>3</td>
<td>I enjoy writing about music, for example on blogs and forums.</td>
</tr>
<tr>
<td>8</td>
<td>I'm intrigued by musical styles I'm not familiar with and want to find out more.</td>
</tr>
<tr>
<td>15</td>
<td>I often read or search the internet for things related to music.</td>
</tr>
<tr>
<td>21</td>
<td>I don't spend much of my disposable income on music.</td>
</tr>
<tr>
<td>24</td>
<td>Music is kind of an addiction for me - I couldn't live without it.</td>
</tr>
<tr>
<td>28</td>
<td>I keep track of new of music that I come across (e.g. new artists or recordings). I have attended _ live music events as an audience member in the past twelve months.</td>
</tr>
<tr>
<td>34</td>
<td>I listen attentively to music for ___ per day.</td>
</tr>
<tr>
<td></td>
<td><strong>Factor 2 - Perceptual Abilities</strong></td>
</tr>
<tr>
<td>5</td>
<td>I am able to judge whether someone is a good singer or not.</td>
</tr>
<tr>
<td>6</td>
<td>I usually know when I'm hearing a song for the first time. I find it difficult to spot mistakes in a performance of a song even if I know the tune.</td>
</tr>
<tr>
<td>11</td>
<td>I can compare and discuss differences between two performances or versions of the same piece of music.</td>
</tr>
<tr>
<td>12</td>
<td>I have trouble recognizing a familiar song when played in a different way or by a different performer.</td>
</tr>
<tr>
<td>18</td>
<td>I can tell when people sing or play out of time with the beat.</td>
</tr>
<tr>
<td>22</td>
<td>I can tell when people sing or play out of tune.</td>
</tr>
<tr>
<td>23</td>
<td>When I sing, I have no idea whether I'm in tune or not.</td>
</tr>
<tr>
<td>26</td>
<td>When I hear a music I can usually identify its genre.</td>
</tr>
<tr>
<td></td>
<td><strong>Factor 3 - Musical Training</strong></td>
</tr>
<tr>
<td>14</td>
<td>I have never been complimented for my talents as a musical performer.</td>
</tr>
</tbody>
</table>
I would not consider myself a musician. I engaged in regular, daily practice of a musical instrument (including voice) for ___ years. At the peak of my interest, I practiced ___ hours per day on my primary instrument. I have had formal training in music theory for ___ years. I have had ___ years of formal training on a musical instrument (including voice) during my lifetime. I can play ___ musical instruments

---

**Factor 4 - Singing Abilities**

4 If somebody starts singing a song I don't know, I can usually join in.
7 I can sing or play music from memory.
10 I am able to hit the right notes when I sing along with a recording.
17 I am not able to sing in harmony when somebody is singing a familiar tune.
25 I don’t like singing in public because I’m afraid that I would sing wrong notes.
29 After hearing a new song two or three times, I can usually sing it by myself.
30 I only need to hear a new tune once and I can sing it back hours later.

---

**Factor 5 - Emotions**

2 I sometimes choose music that can trigger shivers down my spine.
9 Pieces of music rarely evoke emotions for me.
16 I often pick certain music to motivate or excite me.
19 I am able to identify what is special about a given musical piece.
20 I am able to talk about the emotions that a piece of music evokes for me.
31 Music can evoke my memories of past people and places.
<table>
<thead>
<tr>
<th>Please circle the most appropriate category:</th>
<th>1 Complete Disagree</th>
<th>2 Strongly Disagree</th>
<th>3 Disagree</th>
<th>4 Neither Agree nor Disagree</th>
<th>5 Agree</th>
<th>6 Strongly Agree</th>
<th>7 Completely Agree</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. I spend a lot of my free time doing music-related activities.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>2. I sometimes choose music that can trigger shivers down my spine.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>3. I enjoy writing about music, for example on blogs and forums.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>4. If somebody starts singing a song I don’t know, I can usually join in.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>5. I am able to judge whether someone is a good singer or not.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>6. I usually know when I’m hearing a song for the first time.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>7. I can sing or play music from memory.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>8. I’m intrigued by musical styles I’m not familiar with and want to find out more.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>9. Pieces of music rarely evoke emotions for me.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>10. I am able to hit the right notes when I sing along with a recording.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>Question</td>
<td>1 Completely Disagree</td>
<td>2 Strongly Disagree</td>
<td>3 Disagree</td>
<td>4 Neither Agree nor Disagree</td>
<td>5 Agree</td>
<td>6 Strongly Agree</td>
<td>7 Completely Agree</td>
</tr>
<tr>
<td>-------------------------------------------------------------------------</td>
<td>-----------------------</td>
<td>---------------------</td>
<td>------------</td>
<td>-----------------------------</td>
<td>---------</td>
<td>------------------</td>
<td>--------------------</td>
</tr>
<tr>
<td>11. I find it difficult to spot mistakes in a performance of a song even if I know the tune.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>12. I can compare and discuss differences between two performances or versions of the same piece of music.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>13. I have trouble recognizing a familiar song when played in a different way or by a different performer.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>14. I have never been complimented for my talents as a musical performer.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>15. I often read or search the internet for things related to music.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>16. I often pick certain music to motivate or excite me.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>17. I am not able to sing in harmony when somebody is singing a familiar tune.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>18. I can tell when people sing or play out of time with the beat.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>19. I am able to identify what is special about a given musical piece.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>20. I am able to talk about the emotions that a piece of music evokes for me.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>Please circle the most appropriate category:</td>
<td>1 Complete Disagree</td>
<td>2 Strongly Disagree</td>
<td>3 Disagree</td>
<td>4 Neither Agree nor Disagree</td>
<td>5 Agree</td>
<td>6 Strongly Agree</td>
<td>7 Completely Agree</td>
</tr>
<tr>
<td>---------------------------------------------</td>
<td>---------------------</td>
<td>---------------------</td>
<td>------------</td>
<td>-------------------------------</td>
<td>---------</td>
<td>-----------------</td>
<td>-------------------</td>
</tr>
<tr>
<td>21. I don’t spend much of my disposable income on music.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>22. I can tell when people sing or play out of tune.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>23. When I sing, I have no idea whether I’m in tune or not.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>24. Music is kind of an addiction for me - I couldn’t live without it.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>25. I don’t like singing in public because I’m afraid that I would sing wrong notes.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>26. When I hear a piece of music I can usually identify its genre.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>27. I would not consider myself a musician.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>28. I keep track of new music that I come across (e.g. new artists or recordings).</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>29. After hearing a new song two or three times, I can usually sing it by myself.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>30. I only need to hear a new tune once and I can sing it back hours later.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>31. Music can evoke my memories of past people and places.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
</tr>
</tbody>
</table>
Please circle the most appropriate category:

32. I engaged in regular, daily practice of a musical instrument (including voice) for 0 / 1 / 2 / 3 / 4-5 / 6-9 / 10 or more years.

33. At the peak of my interest, I practiced 0 / 0.5 / 1 / 1.5 / 2 / 3-4 / 5 or more hours per day on my primary instrument.

34. I have attended 0 / 1 / 2 / 3 / 4-6 / 7-10 / 11 or more live music events as an audience member in the past twelve months.

35. I have had formal training in music theory for 0 / 0.5 / 1 / 2 / 3 / 4-6 / 7 or more years.

36. I have had 0 / 0.5 / 1 / 2 / 3-5 / 6-9 / 10 or more years of formal training on a musical instrument (including voice) during my lifetime.

37. I can play 0 / 1 / 2 / 3 / 4 / 5 / 6 or more musical instruments.

38. I listen attentively to music for 0-15 min / 15-30 min / 30-60 min / 60-90 min / 2 hrs / 2-3 hrs / 4 hrs or more per day.

39. The instrument I play best (including voice) is ___.
Please tick one of the following:

**Occupational status**
- □ Still at School
- □ At University
- □ In Full-time employment
- □ In Part-time employment
- □ Self-employed
- □ Homemaker/full-time parent
- □ Unemployed
- □ Retired

What is the musical genre you mainly listen to?
(tick only one box)
- □ Rock/Pop
- □ Jazz
- □ Classical Music
What is the Highest educational qualification you have attained?

☐ Did not complete any school qualification
☐ Completed first school qualification at about 16 years (e.g. GCSE/Junior High School)
☐ Completed Second qualification (e.g A levels / High School)
☐ Undergraduate degree or professional qualification
☐ Postgraduate degree
☐ I am still in education

If you are still in education, what is the highest qualification you expect to obtain?

☐ First school qualification (e.g. GCSE / Junior High School)
☐ Post-16 vocational course
☐ Second school qualification (e.g. A-levels / High School)
☐ Undergraduate degree or professional qualification
☐ Postgraduate degree
☐ Not applicable
Appendix B

Residual and Q-Q plots of the LME model in Experiment 1

Figure B1. The residual plot of the LME model in Experiment 1.
Figure B2. The Q-Q plot of the LME model in Experiment 1.
Appendix C

The VIF Table For the Predictors in the LME Model in Experiment 1

<table>
<thead>
<tr>
<th>Predictors</th>
<th>VIF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gold-MSI Active Engagement</td>
<td>1.26</td>
</tr>
<tr>
<td>Gold-MSI Perceptual Abilities</td>
<td>3.68</td>
</tr>
<tr>
<td>Gold-MSI Musical Training</td>
<td>1.24</td>
</tr>
<tr>
<td>Gold-MSI Singing Abilities</td>
<td>2.55</td>
</tr>
<tr>
<td>Gold-MSI Emotions</td>
<td>1.42</td>
</tr>
<tr>
<td>Age</td>
<td>2.00</td>
</tr>
</tbody>
</table>
Appendix D

The VIF Table For a Full Set of Predictors in the GLME Model in Experiment 2

<table>
<thead>
<tr>
<th>Predictors</th>
<th>VIF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gold-MSI Active Engagement</td>
<td>1.56</td>
</tr>
<tr>
<td>Gold-MSI Perceptual Abilities</td>
<td>4.49</td>
</tr>
<tr>
<td>Gold-MSI Musical Training</td>
<td>1.15</td>
</tr>
<tr>
<td>Gold-MSI Singing Abilities</td>
<td>3.01</td>
</tr>
<tr>
<td>Gold-MSI Emotions</td>
<td>2.66</td>
</tr>
<tr>
<td>Age</td>
<td>1.70</td>
</tr>
<tr>
<td>Note</td>
<td>1.00</td>
</tr>
</tbody>
</table>
Appendix E

The Residual Plot of the GLME Model in Experiment 2
Appendix F

The VIF Table For a Full Set of Predictors in the GLME Model in Experiment 3

<table>
<thead>
<tr>
<th>Predictors</th>
<th>VIF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gold-MSI Active Engagement</td>
<td>1.56</td>
</tr>
<tr>
<td>Gold-MSI Perceptual Abilities</td>
<td>6.35</td>
</tr>
<tr>
<td>Gold-MSI Musical Training</td>
<td>1.16</td>
</tr>
<tr>
<td>Gold-MSI Singing Abilities</td>
<td>4.94</td>
</tr>
<tr>
<td>Gold-MSI Emotions</td>
<td>2.83</td>
</tr>
<tr>
<td>Note 1</td>
<td>1.02</td>
</tr>
<tr>
<td>No. of Unique Pitches</td>
<td>1.01</td>
</tr>
<tr>
<td>Decay/Silence 6 Seconds Average</td>
<td>1.92</td>
</tr>
<tr>
<td>Age</td>
<td>2.91</td>
</tr>
</tbody>
</table>
Appendix G

The Residual Plot of the GLME Model in Experiment 3