UNIVERSITY OF SOUTHAMPTON, 2010
FACULTY OF ENGINEERING, SCIENCE AND MATHEMATICS
INSTITUTE OF SOUND AND VIBRATION RESEARCH

# Test-Retest Reliability of the South of England Cochlear Implant Centre (SOECIC) Music Test Battery (MTB): An Investigation of Two Response Methods

Rachel Lamb

Supervised by Rachel van Besouw

A dissertation submitted in partial fulfilment of the requirements for the degree of Master of Science by instructional course.

2010

## Acknowledgements

I would like to thank my supervisor Rachel van Besouw for all her help and support throughout the project.

## Declaration

I, Rachel Lamb, declare that this thesis is my own work, except where acknowledged, and that the research reported in this thesis was conducted in accordance with the principles for the ethical treatment of human subjects as approved for this research by the Ethics Committee at the Institute of Sound and Vibration Research, University of Southampton.

## Abstract

The South of England Cochlear Implant Centre (SOECIC) Music Test Battery (MTB) is a new battery of tests currently under development, designed for the assessment of cochlear implant (CI) users' music perception in the clinic. The MTB consists of pitch discrimination, pitch identification and rhythm discrimination tests. This research project investigated the test-retest reliability of the SOECIC MTB and examined whether two different response methods influenced the test-retest reliability of the tests. Twenty normal hearing participants were recruited to complete the tests using both a three interval three alternative forced choice (3I3AFC) procedure and a three interval two alternative forced choice (3I2AFC) procedure. The test sessions were repeated on three separate occasions, in order to assess test-retest reliability. Eighteen of the twenty participants recruited completed all the test sessions and were included in the data analyses. It was hypothesised that the 3I2AFC procedure would produce more consistent results compared to the 3I3AFC procedure. This is because the presence of a reference tone (3I2AFC procedure) would assist the participant by providing a constant point by which to make a comparison. If the tests were reliable, then the participants' scores would not change significantly when testing was repeated. A comparison of group means showed that scores did not statistically differ for the three test sessions for either response method, indicating that there was not an effect of time. Further statistical analyses showed moderate to strong correlations between the three test sessions when participants were using the 3I3AFC procedure. However, the correlations were weaker (some not statistically significant, $p>0.05$) between the test sessions when participants were using the 3I2AFC procedure. It was concluded that the SOECIC MTB demonstrated test-retest reliability when administered to a small sample of normal hearing listeners. The correlations between test sessions imply that the 3I3AFC procedure should be retained over the 3I2AFC procedure for future testing. It is proposed that further research should focus on investigating whether similar levels of test-retest reliability can be demonstrated when testing CI users.

# Table of Contents

## List of abbreviations

| | |
|---|---|
| **BKB** | Bamford-Kowal-Bench |
| **CAMP** | Clinical Assessment of Musical Perception |
| **CI** | Cochlear Implant |
| **DLF** | Difference Limen for Frequency |
| **FSP** | Fine Structure Processing |
| **HiRes** | HiResolution |
| **JND** | Just Noticeable Difference |
| **MBEA** | Montreal Battery for the Evaluation of Amusia |
| **MTB** | Music Test Battery |
| **MERT** | Musical Excerpt Recognition Test |
| **Mu.S.I.C.** | Musical Sounds in Cochlear Implants |
| **NICE** | National Institute of Clinical Excellence |
| **PMMA** | Primary Measures of Music Audiation |
| **SOECIC** | South of England Cochlear Implant Centre |
| **3I3AFC** | Three interval three alternative forced choice |
| **3I2AFC** | Three interval two alternative forced choice |

# 1 Introduction

This research project aimed to investigate the test-retest reliability of the South of England Cochlear Implant Centre (SOECIC) Music Test Battery (MTB). This is a new battery of tests currently under development, designed for the testing of cochlear implant users in the clinic. In addition, the study investigated whether the response method used, influenced the test-retest reliability of these tests. This dissertation begins with an introduction to cochlear implant technology and then outlines current findings regarding cochlear implant users' speech, rhythm, pitch, melody and timbre perception, making a comparison to normal hearing listeners. The first chapter ends with an examination of a selection of current music tests, culminating in the proposed research question regarding the SOECIC MTB.

## 1.1 Cochlear implants

Cochlear implants (CIs) are a prosthetic device implanted into the inner ear, designed to restore partial hearing to the profoundly deaf. CIs bypass the outer and middle ear mechanisms and transmit information directly to the nerve cells located in the cochlea (Clarke, 2006). CIs are implanted when hair cells in the cochlea are damaged or absent, however neural components connecting the cochlea and auditory nerve are still intact. Electrodes are implanted into the cochlea, which electrically stimulate the auditory nerves. The patient will perceive this stimulation as sound. Successful implantations have resulted in people being able to communicate without lip reading in quiet environments and often show significant improvements at listening in noise.

### 1.1.1 Cochlear implant candidacy

According to the National Institute for Health and Clinical Excellence (NICE) technology appraisal guidance 166 (2009), unilateral CIs are recommended for patients with severe to profound hearing loss in both ears who have demonstrated that they do not receive adequate benefit from acoustic hearing aids. Severe to profound loss is defined as thresholds greater than or equal to 90dBHL. Adequate benefit for acoustical hearing aids is defined for adults as a score of 50% or higher on the Bamford-Kowal-Bench (BKB) sentence test at a level of 70dB SPL. For children, adequate benefit from hearing aids is defined as speech, language and listening skills that are appropriate to the age and developmental stage of the child. Bilateral implantation is recommended for all suitable children, or adults who are blind with other disabilities that increase their need of auditory stimuli for spatial awareness.

## 1.2 The measurement of music perception

A large factor in the study of CI users' performance is that results are often very varied. This is likely to be attributed to the success of different aspects of the surgery as well as the different parameters used by the manufacturers. Electrode depth, input dynamic range, stimulation rate, speech processing strategy and compression function all have different affects on performance depending on the task. Many other factors determined by the CI recipient such as cognitive processing and attitude towards the CI (for example the motivation of the patient) will influence performance. In addition, the aetiology of hearing loss, duration of deafness, and availability of rehabilitation post implantation can influence the outcomes.

The development of music tests is of huge importance in order to assess current abilities, therefore aid in the future development of processing strategies which enhance music perception. They are also likely to be useful in assessing CI candidacy and the assessment of progress to help towards determining in the direction of future rehabilitation. There has been much research examining the music perception of CI users compared to normal hearing listeners. This section will review recent literature which has investigated rhythm, pitch, melody and timbre discrimination and identification. These aspects will be investigated in detail because these are some of the key elements which make up music perception and are likely to be the factors which influence the outcome of music perception. In addition, they are the most relevant components of music perception in terms of the SOECIC MTB. Furthermore, much of the previous research in this area has focussed on these elements of music perception (for example, Gfeller, 1997, 2002; Nimmons et al., 2005).

### 1.2.1 Rhythm perception

The main findings of music perception with CIs have been reviewed by McDermott (2004) and more recently Looi (2008). The literature reviewed indicates that CI listeners tend to perceive rhythm at similar performance levels to listeners with normal hearing. In a study of music perception and psychophysical abilities in CI users, Drennan and Rubinstein (2008) reported that CI users encode rhythm through temporal gaps and amplitude modulations. This kind of information is transmitted successfully, which is confirmed by much research demonstrating CI users' rhythm perception is often close to normal levels of performance. Gelfand (1998) states that normal hearing listeners' temporal discrimination ability has been measured as gap detection thresholds of approximately 2-3 milliseconds.

Kong et al. (2004) carried out an experiment examining the use of temporal cues in three music perception tasks. The tasks involved tempo discrimination, rhythmic pattern identification and melody identification. Four normal hearing listeners and five CI users took part in the tempo discrimination task. The CI users had at least two years of experience with their implant. Participants were given pairs of stimuli which varied in tempo, and they were required to identify the faster tempo. CI users were instructed to listen through their devices at their preferred setting. The results showed that CI users and normal hearing listeners performed similarly when rate discrimination difference limens were compared.

In addition, three CI users from the previous task and four of the normal hearing participants took part in a rhythm pattern identification task. Seven, one-bar rhythmic patterns were presented and participants were required to select the corresponding musical notation from a number of different patterns shown to them on a screen. All participants were trained to read basic musical notation. One CI participant performed close to the performance of the normal hearing listeners; however, the other CI users performed 5-25% worse than normal hearing listeners. It is not possible to conclude from these results that CI users are worse at rhythmic pattern identification, due to the very small sample size. Also, the fact that one CI user did perform close to normal levels suggests that more CI users could perform at similar levels. This type of test would not be practical for the clinic as the whole experiment, including training and testing, took up to five months. This is because participants were required to complete training in basic musical concepts before being able to complete the task. In order for it to be useful for the clinic, the SOECIC MTB has been designed in a way which does not require training prior to testing.

Gfeller et al. (1997) carried out an experiment testing the music perception abilities of 17 CI users who had been implanted with a Cochlear Nucleus device. The results were compared against 35 normal hearing participants. The tests included an adapted version of the Primary Measures of Music Audiation (PMMA) test, the 6-Pulse test, a musical background questionnaire and speech tests. The rhythmic subset of the PMMA contained 14 pairs of short rhythmic patterns. The rhythmic pairs were presented one after the other and the participant had to indicate whether they were the same or different. The 6-Pulse test involved computer generated square waves of the same duration being presented. The gap between each pulse was the same except for one shorter gap which was 10% of the other gaps. The participant had to indicate where in the chain of pulses the shorter gap occurred (towards the beginning or towards the end).

The results of the PMMA rhythmic subset showed CI users' performance to be at levels similar to normal listeners. However, when CI users were given the 6-Pulse task, CI users performed at levels below normal listeners. This demonstrates how a good sensitivity of test is required in order to identify the differences between normal and CI users' rhythm perception ability. It is likely that CI users were not as successful at the 6-Pulse task because it required them to be able to perceive much shorter duration gaps between pulses compared to the rhythm sequences presented in the PMMA rhythm test. The finer gap detection required to complete the 6-Pulse task may have been beyond the limits of the CI or the patients' damaged auditory systems to resolve the fine temporal differences, even if they had been transmitted successfully via the CI. Gfeller et al. (1997) suggested that the difference in performance between the PMMA rhythm subset and the 6-Pulse task may have been partly due to the different requirements of the task. It may have been easier for participants to simply respond with 'same' or 'different' in the PMMA test rather than actually identify the location of the short duration gap in the pulses.

Leal et al. (2003) conducted an experiment testing the music perception of 29 adult CI patients. Tests assessed the participants' ability at pitch, rhythm and timbre discrimination and identification of nursery songs. The rhythm test involved the presentation of 10 pairs of musical pieces. Each piece was separated by 10 seconds of silence. All of the stimuli were presented at the same frequency. Participants were required to respond with whether a pair of items was the 'same' or 'different'. In the identification test the participants were required to state at which point the rhythm of the sequence changed. The results showed that 41% of participants responded correctly to all of the rhythm identification tests and 59% responded correctly to all of the rhythm discrimination tests. Leal et al. (2003) concluded that most participants gained satisfactory scores on all tests; however, it was not possible to know how well they did as no normal hearing participants were tested to make a comparison.

Most of the experiments examining rhythm perception have shown CI users' performance to be close to that of normal hearing participants. However, the experiment by Gfeller et al. (1997) suggested that there are limits to CI listeners' ability to perceive temporal differences. This shows a possible ceiling effect of tests like the rhythm subset of the PMMA. It could be argued that the rhythm tests used in the PMMA are more realistic than the 6-pulse task. This is because the very short latency pulses in the 6-Pulse task are not similar to the type of rhythm component that would be heard in most pieces of music. However, it is important to find out the limits of CI users abilities in order to identify areas for improvement. A reduction in the transmission of temporal cues could have effects on not only music perception but other problematic areas such as speech perception in noise. Therefore, the development of sensitive music tests could provide useful information for a wide range of improvements.

### 1.2.2 Pitch perception

Pitch perception can be broken down into two separate categories: pure tone perception and complex tone perception.

#### Pure tone perception

Processing strategies used in current CIs are unsuccessful at delivering temporal fine structure which is important for hearing pitch differences, differentiating between instruments and appreciating melodies. Gfeller et al. (2002) tested 16 CI users' pitch perception ability and compared it to three normal hearing listeners. The just noticeable difference limen (JND) was measured for 200, 400, 800, 1600 and 3200Hz. Participants were presented with four pure tones (three the same, one different) and the participants were required to indicate which one was different. The results showed that CI users' JNDs were significantly larger than the normal listeners', however there was quite a large amount of variation. Some of the CI users were able to discriminate between intervals slightly narrower than one semitone, while others could not discriminate between intervals an octave apart.

Gfeller et al. (2007) tested 114 CI recipients and compared them to 21 normal hearing listeners. The authors examined 101 CI users who were implanted with long electrode arrays (22mm), as well as 12 users who were implanted with shorter electrode arrays (10mm) and still had the use of some residual acoustical hearing. The task involved participants determining the direction of a pitch change. Gfeller et al. (2007) called this measure pitch ranking ability, which was the probability of a correct response as a function of interval size and base frequency. The stimuli were pure tones ranging between one octave below middle C to two octaves above middle C. These were presented in pairs with pitch differences ranging from one semitone to four semitones. They were also presented in one of three different octave levels (131-262Hz, 262-524Hz and 524-1048Hz), so ability could be compared for different frequency ranges. Three tones were presented to the participants. The first two were the same pitch, whereas the last one was a different pitch. The participants had to respond with whether they thought the last tone was higher or lower than the previous two. The results of this test were also correlated against results for pure tone frequency discrimination tests, familiar melody recognition tasks and speech in noise tasks.

The findings showed that CI users were significantly poorer in pitch discrimination compared to the normal hearing group and the acoustical and electrical stimulation group. The acoustical and electrical stimulation group were worse than the normal listeners, but the largest difference in ability was between the long electrode array group and the acoustical and electrical stimulation group. This demonstrates that low frequency acoustic hearing can be quite significantly advantageous for pitch discrimination compared with traditional electric only CIs. It is likely that the combined electrical and acoustical CI users benefit from the fine structure information available to them through their residual acoustical hearing. This demonstrates the importance of attempting to develop strategies which include more temporal fine structure information.

### Complex tone perception

Gfeller et al. (2002) tested eight normal hearing adults and 46 experienced CI users. Participants were presented with two, one second tones. The tones were from a standard grand piano ranging from 73Hz to 553Hz (three octaves). A two alternative forced choice design was used where the participants were required to indicate whether the second tone was higher or lower than the first one. The minimum interval on this test was one semitone. The results showed that the normal hearing participants had a mean minimum threshold of 1.13 semitones with a range of one to two semitones. This is in comparison with the CI users who had a mean minimum threshold of 7.56 semitones with a range of 1 to 24 semitones. The variability in CI users' performance is clearly apparent and some produced performance near the normal hearing range. Although there were significant differences between the two groups it is possible there may have been an even bigger difference. This is because the normal hearing participants may have been able to discriminate between intervals smaller than one semitone but this was not tested.

Nimmons et al. (2005) used the Clinical Assessment of Music Perception (CAMP) tool to test the music perception ability of eight CI users. There were five men and three women aged between 27 and 76 years. For the pitch test, the frequencies were distributed within the octave surrounding middle C. A two alternative forced choice adaptive procedure was used to determine the threshold. The minimum interval tested was one semitone and the maximum was 12 semitones. The results of the experiment showed that pitch discrimination scores ranged from a pitch discrimination limen of 1 semitone to 9.1 semitones at 185Hz, 11.5 semitones at 262Hz, 9 semitones at 330Hz and 6.5 semitones at 394Hz.

Laneau et al. (2004) tested six CI users' discrimination of musical tones. The stimuli used were tones from five different instruments. These were piano: clarinet, trumpet, guitar and synthetic voice. Three different notes for each instrument were presented with fundamental frequencies of 130.8, 185 and 370 Hz. The frequency intervals for the test notes were one, two, or four semitones from the reference note. The reference note and then the test note were presented and the participant had to respond with whether the second note was higher or lower. The results indicated that most participants were able to discriminate between 2 and 4 semitones. The findings of Nimmons et al. (2005) and Laneau et al. (2004) have shown to be quite variable between CI users but within the range of results found in Gfeller et al.'s (2002) experiment.

### 1.2.3 Melody perception

Kong et al. (2004) tested six CI users and six normal hearing listeners with a melody identification task. Participants were played two sets of 12 familiar songs. One set contained rhythmic as well as melodic information. The other set contained only melodic information; therefore pitch was the only useful cue in identification of the melody. Participants were presented with the songs and then had to select the corresponding title of the song from a list of the 12 titles displayed to them on a screen. CI listeners performed significantly worse in melody identification, in both rhythm and no rhythm conditions. Normal hearing listeners gained scores of 98.3% for the rhythm condition and 97.5% for the non rhythm condition. The CI users scored on average 62.3% for the rhythm condition and 11.7% for the non rhythm condition.

The results suggest that CI listeners appear to rely mainly on temporal (rhythmic) cues in melody identification, as performance was greatly reduced for the non rhythm melodies. This is understandable as experiments testing rhythm perception have shown CI users to perform near normal levels. This is because of the success of current CI devices at transmitting the temporal envelope. Pitch perception tests have shown CI users to perform at levels much lower than normal listeners due to the lack of fine structure information. Therefore, melody perception appears to rely on rhythm cues. However, a problem with this method is that removing the rhythm from the melodies is likely to make the song sound quite different. For example, a song with the same tune but a different rhythm is likely to sound different to the original song. Therefore, it is not possible to be sure that taking the rhythm cues away will leave a perceptually similar sounding melody. It may not be fair to say that they performed worse because they were unable to perceive the melody properly. The perceptual difference in the melody due to removal of the rhythm could have contributed to the CI users' scores being lower in this condition.

Kong et al.'s (2004) experiment shows that in a relatively real world situation, the impairment that CI users have in comparison with normal hearing listeners is quite apparent. However, it does not show us exactly which areas of music perception and what effect this has on their performance.

Gfeller et al. (2005) produced a music test called the Musical Excerpt Recognition Test (MERT) and tested 79 CI users. Gfeller et al. (2005) compared CI recipients with 30 normal listeners' ability to recognise "real world" music excerpts. The MERT contained music excerpts which were 12 to 17 seconds long and were from one of three genres of music. These were classical, country and pop, which were chosen because they were the most commonly heard types of music for the sample. There were 34 target items played to the participant. These were made up of eight familiar excerpts and four obscure excerpts for each genre of music. The excerpts were played to the participants and then they had to indicate if they recognised the song or not. If they indicated that they did, they were required to answer a number of questions about the song to show that they knew what it was.

The results showed that CI users were significantly poorer at recognition of familiar excerpts compared to the normal hearing listeners. CI implant users were most successful at recognising excerpts from the country genre, then pop excerpts and least successful at excerpts from the classical genre. This is likely to be due to the higher number of rhythm and speech cues present in country and pop music compared to classical music. These results are consistent with what is known about the current speech processors which transmit the envelope of the sound relatively successfully, however transmit very little fine structure information important for pitch perception.

A problem with using real world song excerpts is that the participants' performance will be greatly influenced by their ability to recall the title and/or composer of the song. This does not necessarily demonstrate their true song recognition ability. It also does not provide information of exactly which parts of the music sample they are hearing. The fact that three different genres were used gave some indication that CI users were more successful at recognising music with more rhythm and speech components. However, a music perception test such as the SOECIC MTB which tests each element of music in isolation would be more useful to identify exactly how well CI users are performing. Better performance at individual music component tests would be likely to correlate with ability to recognise music excerpts similar to the ones used in Gfeller et al.'s (2005) experiment.

### 1.2.4 Timbre perception

McDermott (2004) described timbre as differences which are apparent when a note of the same pitch and loudness is played on different instruments. Most studies on the perception of timbre have tested participants' ability to discriminate between sounds from different instruments. Gfeller et al. (2002) tested 51 CI users and compared them to 20 normal listeners. The stimuli consisted of recordings of eight different instruments playing the same sequence of notes. The participants were required to select the correct instrument from a choice of 16 different options. The result showed that the normal participants got an average score of 90.9% correct, whereas the CI users scored an average of 46.6% correct.

Leal et al. (2003) tested timbre perception by presenting 20 CI participants with short melodies played by instruments from three different instrumental families. These were a melody played on the trombone (wind family), a melody played on the piano (percussion family) and a melody played on the violin (string family). The participants were asked to identify each instrument. The results showed that 69% of participants correctly identified all three instruments. A limitation of this study is that although only three instruments were tested, people who were unfamiliar with different types of instruments may have struggled to name them correctly. Gfeller et al. (2002) gave a selection of choices for the participants from which to choose their response. However, Leal et al.'s (2003) study had so few stimuli this was not really possible as participants could have guessed.

*Table 1: Summary of normal and CI hearing*

| | **Normal hearing** | **CI hearing** |
|---|---|---|
| **Dynamic range** | 120dB | Input 60dB, Outcome 5-15dB (variable depending on patient) |
| **Temporal discrimination** | Gap detection threshold approx 2-3 milliseconds | Near normal limits for standard musical timing however impaired at faster rates. |
| **Pitch range** | 20-22000 Hz | 100-8000Hz (Implant range) |
| **Timbre perception** | Can distinguish between instruments. | Can distinguish between some instruments, however often interfamily confusions |
| **Complex tone discrimination** | Can usually distinguish between tones of less than 1 semitone apart. | Variable performance ranging between 2 semitones to more than 1 octave. |
| **Frequency discrimination** | DLF Ranging from 1 Hz at 200Hz, 2-3Hz at 1000-2000Hz, 68Hz at 8000Hz | Variable: DLF Ranging between 4Hz to 200Hz for a 200Hz tone. |
| **Perceptual frequency bands** | Approx. 24 distinct critical bands that are perceptually separated. | May have up to 22 electrodes but these are unlikely to stimulate distinct pitch percepts. Therefore, more likely to be around 8 distinct pitch percepts, very variable depending on the patient. |

### 1.2.5 Preference for music

Table 1 above gives a summary of normal hearing compared to CI hearing. It is clear that there are significant differences which are likely to influence a CI user’s enjoyment of music. Leal et al. (2003) gave CI users a questionnaire to assess their interest in listening to music before and after implantation and how much time they spent listening to music. The results showed that 86% of participants had less interest in listening to music after implantation and 38% of participants reported that they did not enjoy listening to music. However, 21% reported they enjoyed listening to music and searched for opportunities to listen, for example, at concerts and shows. This shows that in general music perception is worse for CI users; however, some do gain enjoyment from music. The fact that some users do enjoy music with current devices, demonstrates the potential benefit CI users could get if improvements were made to implants in terms of music perception.

Looi and She (2010) developed and administered a questionnaire to assess CI recipients’ rating of music after implantation and what aspects of the listening experience contribute to it being less enjoyable. The results of the questionnaire showed that after implantation respondents generally found listening to music less enjoyable and also stated that music did not sound how they would expect it to sound. However, it could be improved by controlling the environment in which the music was listened to, choosing certain types of music and using a contralateral hearing aid where possible. The results of the questionnaire were used to help develop a music training programme which may improve CI users' experience of listening to music. In Gfeller and Lansing's (1992) experiment examining the PMMA tests, a strong positive correlation between music listening habits after implantation and tonal accuracy was found. This shows how experience and exposure to music over time could improve music perception ability post implantation, thus training has the potential to maximise the enjoyment of music. However, it is also possible that participants with better tonal accuracy were more likely to listen to music. Therefore, it may not have been the exposure to music which improved tonal accuracy. Although it is not possible to determine the direction of causation from this correlation, Looi and She's (2010) questionnaire indicates that if CI users can be helped in how they listen to music, this may improve their experience of music.

## 1.3 Music tests

Research has indicated that CI users see music perception as an important element of their auditory lives and the ability to listen to music is a strong positive quality of life factor (Gfeller et al., 2000). The development of music tests is very important in order to assess performance and help towards the future development of CI technology so improvements can be made to music perception. They are also likely to be useful in CI candidacy, the assessment of progress and to help in decision making regarding future rehabilitation. This section describes in more detail, some of the currently used music perception tests.

### 1.3.1 Primary Measures of Music Audiation (PMMA)

Gordon (1979) developed the PMMA which is a standardised test designed to assess listeners' ability to distinguish differences in short tonal and rhythmic patterns. This is made up of 40 rhythmic or tonal pairs of patterns. The first item of the pair is presented and 1.5 seconds later the second item is presented. Each pair is presented five seconds apart. The tonal subset consisted of melodic patterns which ranged from two to five notes (260Hz to 694Hz). In the 'same' condition, the pairs of tones were identical. However, the 'different' pairs differed in a frequency range of one or two notes. For the rhythm subset, patterns were all of the same frequency (520Hz), the differences in the notes were in duration or intensity. Participants are required to respond with whether the pairs were the 'same' or 'different'. A percentage correct score is calculated for the tonal and rhythm subset scores for each participant. Gfeller and Lansing (1992) reviewed the effectiveness of the PMMA for testing CI users and found it to be a useful tool. However, as discussed earlier the rhythm subset may not be sensitive enough to identify finer differences between normal listeners and CI users. A test similar to the 6-pulse test could be an alternative to test shorter temporal intervals (Gfeller et al., 1997).

### 1.3.2 Clinical Assessment of Musical Perception (CAMP)

Nimmons et al. (2008) stated that many of the earlier music tests are not standardised tests, so cannot be directly compared against each other. Nimmons et al. (2008) designed a short computerised test called the Clinical Assessment of Music Perception (CAMP). This included tests for pitch direction discrimination, melody identification and timbre identification. The pitch direction discrimination test used digitally synthesised complex tones which were created from a recording of a piano middle C (262Hz). A two alternative forced choice paradigm was used, with a one up one down adaptive procedure. For each trial a reference tone at the fundamental frequency, and a higher pitched tone, were played in a random order. The difference between the reference tone and the higher tone reduced with the adaptive procedure. The smallest difference was one semitone and the largest was one octave. Threshold values were calculated from the mean of the last six reversals, over three separate trials.

The melody identification test involved participants listening to 12 melodies and choosing the title from a closed set questionnaire if they recognised it. For the timbre test, participants were presented with stimuli played by eight different instruments. All of the instruments played the same melody at the same tempo. Participants were played each instrument sample three times and then they were required to identify the instrument from a closed set. Results reported in section 1.2.2 have shown the CAMP to have good sensitivity. Nimmons et al. (2008) concluded that the CAMP tests are capable of showing a broad spectrum of abilities of music perception. A limitation of the melody recognition task is that several of the listeners found it very difficult and reported they did not know the melodies well enough in the first place. This needs to be addressed if a standardised test is going to be used universally in clinics.

Kang et al (2009) tested the validity and test-retest reliability of the CAMP on 42 CI users and 10 normal hearing listeners. Pitch direction discrimination, melody recognition and timbre recognition were tested and compared against consonant-nucleus-consonant (CNC) word recognition scores and spondee recognition thresholds. To assess test-retest reliability, the testing was repeated a few days later for CI users. Pitch tests used an adaptive procedure to measure just noticeable difference limens in a range between 1 and 12 semitones. Melody and timbre tests involved the recognition of 12 commonly known songs and eight musical instruments.

Statistical analysis involved assessment of whether the music tests correlated with the speech test data as well as whether the music test results from the first session of testing correlated with the music test results from the second session of testing. The results showed that all the music tests correlated significantly with the speech tests and there was a moderate to strong correlation between testing session one and two. This shows good test-retest reliability of the CAMP tests. A limitation of these tests however, is that the smallest interval tested is one semitone. It is clear that normal hearing listeners can distinguish between differences smaller than one semitone and it is possible that CI users may be able to in the future. Therefore it is important to avoid ceiling effects in this type of test. The SOECIC MTB is measured in cents (there are 100 cents in a semitone) so much smaller just noticeable difference limens can be measured.

### 1.3.3 Med-EL Musical Sounds in Cochlear Implants (Mu.S.I.C) Test

The Med-EL Mu.S.I.C. Test was developed by Fitzgerald et al (2006). It is a battery of six objective subtests which assess pitch, rhythm, melody, harmony and timbre perception and two subjective subsets which assess emotion and dissonance perception. The pitch tests involve presenting the participant with pairs of tones, one tone being higher than the other. Participants are required to indicate whether the first or second tone was higher. A problem with this type of test is that it combines both pitch discrimination and pitch identification ability. These are two separate skills which could be independent of each other, for example, someone may be able to tell if two tones are different in pitch but not identify which was higher than the other. The rhythm subset uses a same-different paradigm where the participant is presented with pairs of rhythms and has to respond with whether they are the same or different. The melody test uses the same format as the rhythm tests.

### 1.3.4 Montreal Battery for the Evaluation of Amusia (MBEA)

The MBEA was developed by Paretz et al. (2003) initially as a test to identify congenital amusia. However, more recently it has been used by Cooper et al. (2008) to examine music perception ability in normal hearing and CI listeners. Paretz et al. (2003) have shown this test battery to have good sensitivity, reliability and validity based on the testing of 160 normal hearing listeners. This battery of tests consists of six subsets which measure scale, contour, interval, rhythm, meter and melody memory. Most of the tests included in the MBEA implement a same-different paradigm. The scale, contour and interval tests assess different components of pitch perception ability through the presentation of slightly altered melodies. Participants are required to indicate if the pairs of stimuli are the same or different. The rhythm subset follows a similar format where pairs of melodies, some with alterations to the rhythm are presented, and participants have to indicate whether the pairs are the same or different. The meter tests assess the temporal component of music perception in another way. It requires participants to identify whether the melodies presented are either a ‘march’ or a ‘waltz’. Finally, the melody memory test involves playing 15 of the melodies which will have been played previously during the rest of the testing and 15 new melodies. The participant is required to respond with ‘yes’ or ‘no’ to indicate whether they recognised the melody as one which had been played earlier.

Cooper et al. (2008) demonstrated that the MBEA produced similar results to other studies which have tested normal hearing and CI users. However, the authors noted that several CI users were responding at chance level, suggesting that some of the tests were too difficult. This is something that needs to be improved in order to make the tests useful for the assessment of CI users. An advantage of the MBEA tests is that they assess several aspects of the main music perception components. For example pitch perception is assessed via three different tests (scale, interval and contour) and temporal aspects are assessed via two tests (rhythm and meter). Each test lasts approximately 10 minutes each so a large amount of information can be gained in a short space of time.

### 1.3.5 Other custom tests

Kong et al. (2004) implemented a selection of tests to assess tempo discrimination, rhythmic pattern identification and melody identification. The stimuli used in the tempo discrimination task consisted of a one bar pattern of beats presented in a 4/4 time signature. The rhythm patterns were presented using four standard tempos, these were: 60, 80, 100, and 120 beats per minute. Two different tempos were presented one after the other in a pair and the participant was required to identify the faster tempo. A two interval forced choice paradigm was used. The threshold was calculated as the beats per minute tempo which produced a 75% correct score.

The stimuli used in the rhythmic pattern identification test consisted of seven one bar rhythmic patterns. Participants were presented with one 4/4 bar containing four equally separated notes, then they were presented with another bar with the same notes, apart from the second note was varied using eighth and sixteenth notes. Participants were required to choose the musical notation which matched the rhythmic pattern they had heard. This was tested at four different tempos (60, 90, 120 and 150 beats per minute).

The melody identification test consisted of two sets of familiar songs. Each song consisted of 12 to 14 notes and played within a frequency range of 207Hz to 523Hz. In one condition the melodies contained both rhythm and melody cues, and in the other condition the rhythm information was removed. Participants were required to select the title of the song presented from a closed set of answers displayed on a computer screen. The music tests designed by Kong et al. (2004) have produced results which are consistent with other music perception tests; however, some of the tests required training in basic musical notation which is not practical in a clinical setting.

## 1.4 Music perception test methodology

The PMMA and Kong et al.’s (2004) tests have used an adaptive two alternative forced choice procedure (for example a ‘same’ or ‘different’ response to two stimuli). Leek (2001) compared two alternative forced choice methods and three alternative forced choice methods and found that three alternative forced choice procedures tend to produce more reliable and consistent thresholds. During preliminary music perception testing using the SOECIC MTB, it became apparent that the 3I3AFC procedure could be difficult in terms of concentration for the participant. This was perhaps because participants needed to pay attention to all three intervals before choosing their response. The SOECIC MTB has been designed to present stimuli at a relatively slow rate compared to other frequency discrimination tests. This was because initial testing of the music tests showed that CI users required a gap duration of approximately at least one second in order to hear each sound effectively. It is possible that the slow pace of the test could introduce a cognitive factor to the test, where it becomes more of a challenge to recall all three intervals. Therefore, it was proposed that a 3I2AFC procedure may produce more reliable results with a test of this nature.

A 3I2AFC procedure could be implemented in a way that the first interval would be a reference tone, and the participant would have to indicate with whether the second or third interval was different to the reference tone. The difference between the 3I2AFC procedure compared to a 3I3AFC procedure is that participants would be more likely to choose the correct response when choosing between two intervals compared to three. Therefore, one would expect that participants would achieve smaller discrimination scores using the 3I2AFC compared to the 3I3AFC procedure. However, there is a possibility that participant's results may show an advantage for the 3I2AFC procedure over and above the increased level of chance. The presence of the reference tone could assist the participant by providing a consistent point by which to make a comparison. This is something that is not present using the 3I3AFC procedure because the different tone could be any one of the three intervals.

The CAMP test is designed to use a one up one down adaptive procedure. This type of method is based on a 50% correct responding level. Levitt (1971) suggested the use of the transformed up-down method which uses an unequal number of up and down steps, such as two down one up. This is believed to be a better estimate of threshold as it is based on a 71% correct responding level. Other configurations such as three down one up methods can be used but are more time consuming. Leek (2001) explained how adaptive measures are an efficient method for obtaining thresholds compared to non-adaptive testing. They also show relatively high levels of accuracy, stability and reliability. The SOECIC MTB uses a two down one up (correct responding level of 71%) method which appears to be a suitable compromise between an improvement on a one up one down procedure (correct responding level of 50%) and ensuring the test does not become too lengthy (which may be the case for three down one up method with a correct responding level of 79%).

During melody and timbre tests (for example using CAMP), often participants are presented with a closed set of options from which they must choose. It is questionable if this is the most valid way of testing melody recognition as offering them the correct answer may help them recall a familiar melody. However, the other option is to ask participants to recall the name of the melody using an open set paradigm.

It is clear that there are a number of problems in the testing of melody and timbre ability. These are tests of the ability to recognise stimuli taken from real life situations, which in many cases is important because it shows how a participant is coping with day to day situations. However, melody and timbre are directly influenced by pitch and rhythm perception ability. Therefore, it is perhaps more useful to assess pitch and rhythm ability in order to gain information for the improvement of devices and processing strategies. It could be argued, however, that pitch discrimination tasks do not necessarily indicate whether a CI user's ability to perceive a pitch difference is 'musically' useful. For example, a CI user may be able to perceive a difference, but they may not be able to identify the direction or may even perceive it in the opposite direction. This means that pitch discrimination scores may not be directly related to a CI user's ability to recognise and/or appreciate melodies. The development of pitch identification tests may help in determining if the CI user at least perceives the pitch differences in the correct direction. Therefore, a tool combining these two scores could be useful in providing a slightly more informative pitch perception measure.

## 1.5 SOECIC Music Test Battery (MTB)

The SOECIC Music Test Battery (MTB) is a series of tests currently being developed at the South of England Cochlear Implant Centre (SOECIC). This is a computer based set of tests for pitch discrimination and identification and rhythm discrimination used at the SOECIC for preliminary CI investigation of music perception.

### 1.5.1 Test stimuli and procedure

Pure tone (sine) or complex tone (piano) stimuli can be presented at high (880Hz to 1397Hz) or low frequency (220Hz to 349.2Hz) ranges. For the discrimination tests, three tones are presented. Two of the tones are the same and one is of a higher or lower pitch. Participants are required to select the tone which is different. This uses a three interval three alternative forced choice procedure (3I3AFC). For the identification part of the test, the participant is required to respond with whether the different tone is higher or lower, implementing a two alternative forced choice procedure.

The rhythm test consists of three snare drum beats being presented one after the other. Two of the beats have the same interval between them, and one is presented a certain number of semi-quavers early or late. The level of difficulty changes adaptively with correct or incorrect answers. The adaptive procedure uses a three alternative forced choice, two down one up method. The number of reversals can be set by the tester.

The SOECIC MTB has previously been used to investigate the changes in CI users' performance after upgrading from the Tempo+ CI to the Opus 2 CI (both manufactured by Med-EL), (Paynter, 2010). It has also been used to compare normal hearing listeners' music perception ability with and without prior musical training (Paynter, 2010). There is currently very little information on the test-retest reliability of the SOECIC MTB. Research into this area could help identify areas for improvement, for example, indicating possible alterations to parameters which might enhance reliability of the test.

## 1.6 Summary and research question

The present review has examined research investigating the current successes and limitations of CIs in terms of music perception and compared this to normal hearing listeners. In addition, it has investigated how music perception is tested and examined the constraints of these tests and how they could be improved. The main perceptual findings reflect the limits in current speech processing strategies and surgical techniques. Perception of rhythm has been shown to be close to normal limits for simple tests of rhythm discrimination (Kong et al., 2004). More challenging tests such as the 6-pulse task have shown that CI users have not got rhythm discrimination abilities matching normal hearing participants (Gfeller et al., 2002). The success of CI users at some of the simple rhythm discrimination tests is consistent with the relatively intact transmission of the temporal envelope through present speech processing strategies. However, lower scores at the 6-pulse task show improvements could still be made in this area.

Less success has been found in terms of pitch perception. Almost all the literature reviewed indicates that performance of CI users at pitch and melody perception is significantly lower than normal listeners, although performance is hugely variable. CI users' frequency difference limens can range from one semitone to more than an octave. In melody identification tasks, participants were much more successful at identifying melodies with a strong rhythmic component compared to melodies with rhythm cues removed (Kong et al., 2004). The reduction in pitch perception ability is predominately due to the lack of temporal fine structure information transmitted through current speech processing strategies. The fact that CI users were more successful at identifying melodies with a strong rhythmic component reflects the better rhythm perception abilities. Strategies (FSP and HiRes) have attempted to incorporate some temporal fine structure information; however, there has been little research to show any significant improvements.

In experiments examining CI users' enjoyment of music, the majority of participants have reported reduced enjoyment of music post implantation and a significant number report that music sounds unpleasant (Leal et al., 2003). Further advances in the incorporation of fine structure information are really important in order to improve music perception. Music tests are useful in determining the ability of current CI users for providing guidance for future developments. They can also be used as indicators for CI candidacy by testing pitch and rhythm perception prior to implantation. Another use of music tests is that they can be used to evaluate new developments in devices and processing strategies through simulations before or after implantation.

The SOECIC MTB has been developed as a simple, easy to use tool designed to be suitable for use in the clinic. A measurement tool must be valid and reliable if it is to produce true and accurate results. The validity of a test refers to whether the test is actually measuring what it is described to measure, i.e. the results are not affected by any confounding variables. The reliability of a test refers to whether a set of results obtained on one occasion are the same when the test is carried out on another occasion. Test-retest reliability is an important factor and needs to be investigated in terms of the SOECIC MTB. It is crucial that the SOECIC MTB is a reliable measuring tool in order for it to be used in a clinical setting. This is because results need to be consistent in order for comparisons to be made between findings. Also, if it is to be used for CI candidacy, evaluation of progress for determining rehabilitation and evaluation of new developments, the MTB must prove to produce consistent results so that meaningful conclusions can be drawn.

In this research project the test-retest reliability of the SOECIC MTB was investigated. This involved the testing of normal hearing listeners on three separate occasions. The fact that the SOECIC MTB is designed for CI users means that it would be useful to examine the test-retest reliability on the CI population. However, normal hearing listeners were tested as previous music perception testing has shown them to be a less variable population compared to cochlear implant users (see section 1.2). This was important in order to obtain an accurate indication of the basic test-retest reliability of the SOECIC MTB. Also, there were a lot more normal hearing listeners available for recruitment which increased the potential for a larger sample. Participants were tested three times as this was designed to help in determining between any practice effects and test-retest reliability. It is possible that participants could have improved for the first and second session but further improvement may have levelled off. This means that test-retest reliability could still be assessed even if some practice effects were present.

Musicianship was controlled as it has previously been suggested that this can have an influence on music perception ability. Paynter (2010) found that musicians attained significantly smaller pitch and rhythm discrimination scores compared to non-musicians. Cooper et al. (2008) stated how further research into music perception in normal hearing participants and CI users, must take musicianship into account.

The proposed research question was: Do the tests currently available in the SOECIC MTB exhibit test-retest reliability? In addition, two different response selection methods were tested to assess if this influenced test-retest reliability. A three interval three alternative forced choice (3I3AFC) procedure was compared to a three interval two alternative forced choice (3I2AFC) procedure. Thus, the second research question was: Does a 3I3AFC procedure or a 3I2AFC procedure show more test-retest reliability? If the tests were reliable, then the scores of participants taking the tests would not change significantly when testing was repeated. Also, discrimination and identification scores from repeated test sessions would correlate significantly. If the scores were different between test sessions, there may be evidence that there were practice effects or there was a lack of test-retest reliability.

# 2 Method

Prior to experimentation, ethical approval was sought and granted from the University of Southampton Institute of Sound and Vibration Research Human Experimentation Safety and Ethics Committee.

## 2.1 Participant recruitment and screening

Participants aged between 20 and 30 years (Mean age = 24.4, SD = 2.4) were recruited opportunistically from the University of Southampton campus and surrounding area. Screening involved Otoscopy and Pure Tone Audiometry (PTA) to identify normal hearing listeners. This was carried out according to the British Society of Audiology (2004) recommended procedure. Participants were selected for testing if PTA thresholds were ≤20dBHL at 1000, 2000, 4000, 8000, 500 and 250Hz.

Participants were also given a questionnaire regarding musicianship (see appendix 2). Participants were excluded if they reported that they had gained at least grade five standard in an instrument or voice, and currently or within the past five years, had taken part in musical activity. This selection criterion was chosen based on the criteria previously used by Paynter (2010) for the selection of musicians and non musicians. There were 21 participants recruited, although only 20 completed the full experiment (5 males and 15 females). One participant did not continue testing because they were unable to complete the task at the least difficult level.

## 2.2 Test set-up

The experiment was carried out in a sound-proofed booth measuring 4 metres by 4.5 metres. Figure 1 shows the layout of the experiment, which was designed to replicate the layout used by Paynter (2009). A Behringer Truth B2031A Loudspeaker was positioned one meter in front of the participant's chair. This was connected to an Edirol UA-1X sound card which was connected to the experimenter's laptop computer. A sound screen was placed behind the participant's chair in order to help prevent standing waves, which may have occurred during presentation of the sine tone stimuli. Standing waves needed to be avoided as they may have interfered with the participants' ability to discriminate between the sine tone stimuli. A Dell flat screen monitor was placed at a comforTable viewing distance to the right hand side of the participant. The monitor was connected to the experimenter's laptop computer, which was positioned on a Table on the other side of the sound screen, out of view of the participant. The SOECIC MTB response screen was displayed on the monitor. The participant was required to make their response by clicking on the appropriate box displayed on the monitor using a mouse held on their lap. The experimenter stayed in the booth throughout the testing, controlling the presentation of the conditions via the laptop computer.

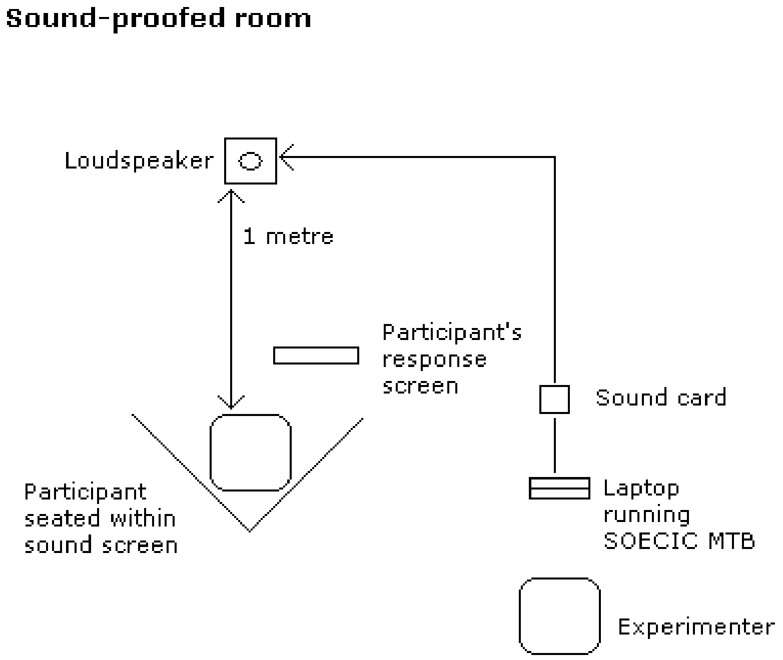
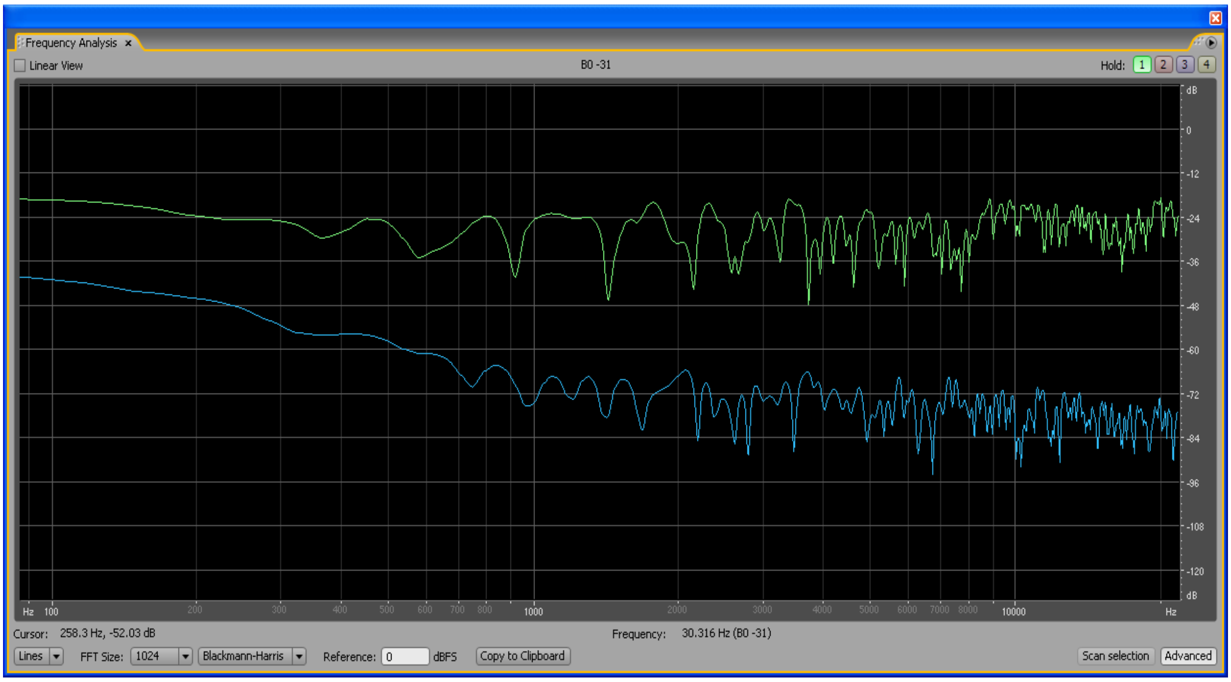

Sound-proofed room
Loudspeaker
1 metre
Participant's response screen
Sound card
Participant seated within sound screen
Laptop running SOECIC MTB
Experimenter

*Figure 1: Layout of the experiment.*

## 2.3 Stimuli and calibration

Before testing, ambient noise levels were measured using a Kamplex sound level meter ensuring that levels did not exceed 30dB(A). The output of the loudspeaker was measured using a Bruel and Kjaer (Type 2230) Sound Level Meter. This was carried out by generating white noise in adobe audition and then playing it through the loudspeaker at 60dB(A). This was then recorded using the Bruel and Kjaer (Type 2230) Sound Level Meter. The output waveform can be seen in Figure 2, the upper line represents the waveform of the original white noise and the lower line shows the frequency analysis of the output of the loudspeaker. Ideally, this experiment should use a loudspeaker with a flat frequency response. This is because the stimuli for pitch tests should not be altered when being presented to the participant. However, this is generally not possible, particularly in a non anechoic room.

Figure 2 shows how the frequency response of the loudspeaker had some differences from the original waveform. The upper trace has a flatter response (particularly in the low frequency region below 600 Hz). It is possible that the boost in low frequency response may not have been due to the loudspeaker alone. The test room was a sound-proofed booth however it was not anechoic. There were windows and cupboards with reflective surfaces. This means that some of the low frequency boost could have been due to the response of the room. This was not viewed as a significant problem as the room selected is similar to the type of room that would be used in clinic. In general, clinic rooms are not anechoic therefore this set up is representative of how the SOECIC MTB software is likely to be utilised.



Cursor: 258.3 Hz, -52.03 dB

Frequency: 30.316 Hz (B0 -31)

FFT Size: 1024 Blackmann-Harris Reference: 0 dBFS

*Figure 2: Frequency analysis of generated white noise (upper line) and frequency analysis of white noise played via the Behringer Truth loudspeaker and recorded by a sound level meter (lower line).*

Participants were presented with three different types of stimuli via the loudspeaker. These included sine tones and piano tones presented at a master level of 60dB(A) and side-stick drum beats at a level of 55dB(A). Before testing, the stimuli output was calibrated using a Kamplex sound pressure level meter. The master volume on the laptop computer was kept constant and then levels on the SOECIC MTB software were adjusted for the three different types of stimuli. In order to eliminate the possibility of loudness cues interfering with the participants pitch discrimination scores, a roving volume setting was implemented for the pitch tests. For the roving volume setting, maximum (63dB(A)), master (60dB(A)) and minimum (57dB(A)) levels were adjusted through the SOECIC MTB software. These settings needed to be adjusted every time a condition using a different stimulus type was carried out.

## 2.4 Procedure

Participants were seated in the sound proofed booth ensuring that they were able to comfortably see the monitor displaying the response screen. The nature of the experiment was explained and participants consented to taking part in the experiment (see appendix 1). Participants were instructed to read a set of written instructions. The written instructions were comprised of two sheets; one detailing the tests using the three interval three alternative forced choice (3I3AFC) procedure (see appendix 3) and the other explaining the three interval two alternative forced choice (3I2AFC) procedure (see appendix 4). Participants were also given a verbal explanation of the tests and provided with the opportunity to ask any questions. Participants were then given a practice session to ensure they understood the difference between the two different procedures.

Participants were presented with six different conditions in total. These included:

1. Sine tone discrimination and identification- 3I3AFC
2. Sine tone discrimination and identification- 3I2AFC
3. Piano tone discrimination and identification- 3I3AFC
4. Piano tone discrimination and identification- 3I2AFC
5. Rhythm discrimination- 3I3AFC
6. Rhythm discrimination- 3I2AFC

The order of conditions for each participant was randomised using a Latin square. Participants were tested on two more occasions where the experimental procedure was repeated. The second session was carried out on average two days after the first session and the third session two days after the second session. It was important to try and test the participants at the same time of day for each session because this could have had an influence on the participants’ performance at the tests. For example, the same participant carrying out the tasks early in the morning on one day and mid-afternoon on another day, may have performed differently due to how they were feeling at that time of day. Therefore, the experimenter aimed to test participants at the same time of day for each session but this was not always possible. Participants were given a break of a maximum of five minutes between each condition.

For the 3I3AFC pitch discrimination tests participants were presented with three sine tones or three piano tones ranging from 220Hz to 1397Hz, presented at a level of 60dB(A). Each tone was one second in length and separated by one second of silence. One of the tones was a different frequency to the other two tones. Participants were instructed to click on the box which referred to the tone that was different. The pitch identification test followed on from the pitch discrimination tests. The participant was instructed to indicate whether the tone that was different, was higher or lower than the other two tones. For the 3I2AFC procedure the first tone was a reference tone and the participant had to respond with whether the second or third tone was different to the reference tone. The identification test followed on from this, where participants had to indicate whether the different tone was higher or lower than the reference tone.

For the 3I3AFC rhythm discrimination tests, participants were presented with three phrases of three side-stick drum beats, presented at a level of 55dB(A). Each phrase was separated by one second of silence. The beats were equally spaced for two of the phrases and one of the phrases contained three beats with one of the beats occurring slightly early or late. Participants were instructed to click on the box which referred to the phrase that was different. For the 3I2AFC tests, the first phrase was a reference phrase which was always three equally spaced beats and participants had to indicate whether the second or third phrase was different to the reference phrase.

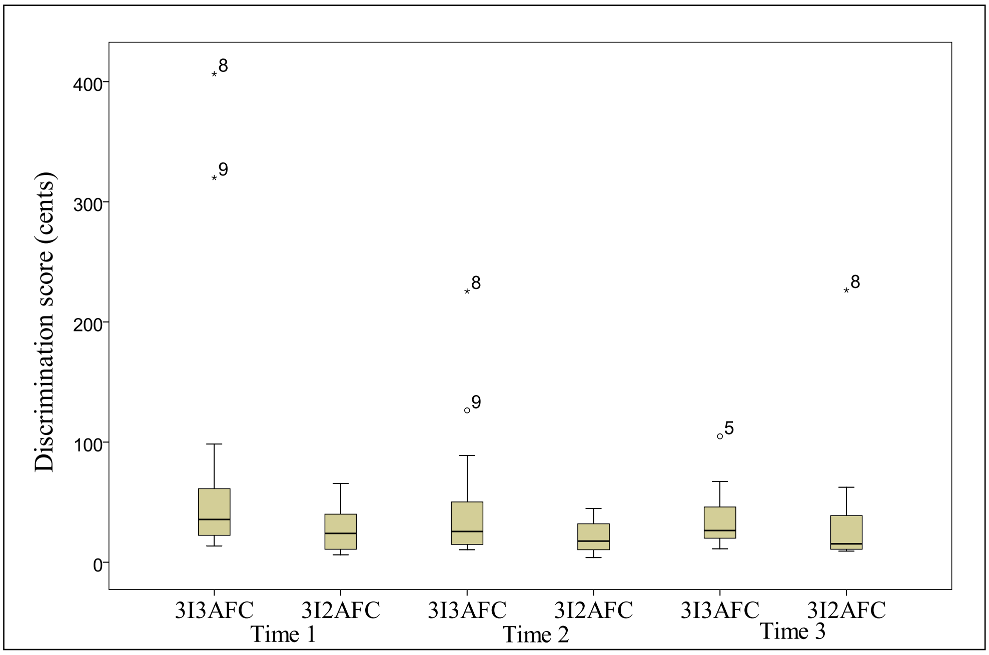
All of the tests used an adaptive two down, one up method. The software continued presentation until seven reversals had been obtained. Threshold measurements were calculated by averaging the discrimination scores over the last five reversals (excluding the first two reversals). The first two reversals were excluded as it normally takes at least two reversals to get down to threshold from the easier starting point of the tests. Identification scores were obtained by calculating the percentage of correct pitch identifications in the last five reversals. Each test session took on average 50 minutes.

# 3 Results

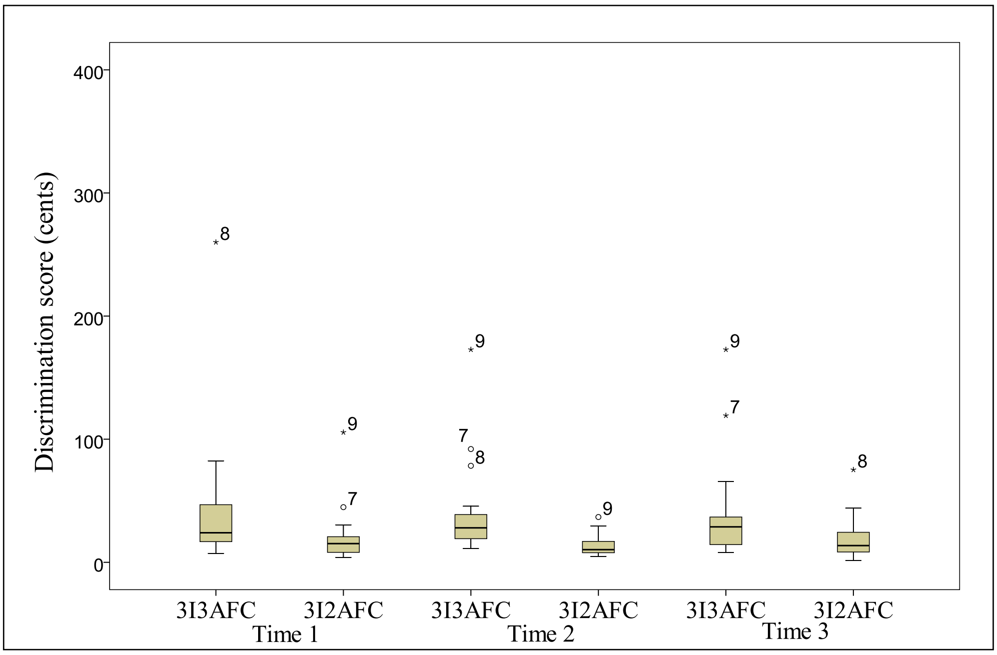
The first section of results (section 3.1) examined participants’ pitch discrimination scores for the sine tone and piano tone tests at time 1, time 2 and time 3. This included investigating scores obtained using the 3I3AFC procedure compared to scores obtained using the 3I2FAC procedure. Section 3.2 examined pitch identification scores for the two procedure types at time 1, time 2 and time 3. Section 3.3 and 3.4 compared the pitch discrimination and identification scores from the present paper with data collected by Paynter (2010). Section 3.5 examined rhythm discrimination scores for the 3I3AFC procedure compared to the 3I2AFC procedure at time 1, time 2 and time 3. Finally, section 3.6 compared the rhythm discrimination scores obtained in the present paper with Paynter’s (2010) results. See Appendix 5 for a CD-ROM containing the experimental data.

## 3.1 Pitch discrimination

The data for the piano tone and sine tone pitch tests using the two different procedures for time 1, time 2 and time 3 is shown in the box plots in Figure 3. It is clear that there were some extreme outliers (participants 8 and 9). Kalmus and Fry (1980) examined the prevalence of congenital amusia (or ‘tone deafness’) in the general population and reported a Figure of approximately 4%. This suggests that perhaps some of the outliers in the data could be due to participants exhibiting some level of congenital amusia. However, the fact that on some occasions participants 8 and 9 performed at similar levels to the other participants but at other times performed significantly worse, suggests that they may have not been concentrating on the task to an accepTable level. The decision was made to remove any participant performing at level above 200 cents as past research indicates that most normal hearing participants can discriminate less than one semitone (Gfeller et al., 2002, findings cited in section 1.2.2). Figure 4 shows the data with participants 8 and 9 removed. The box plots show some new outliers with the adjusted scale, however, there are no participants obtaining discrimination scores at levels above 120 cents which is more representative of normal hearing average pitch discrimination scores found in previous research (for example, Gfeller, 2002).

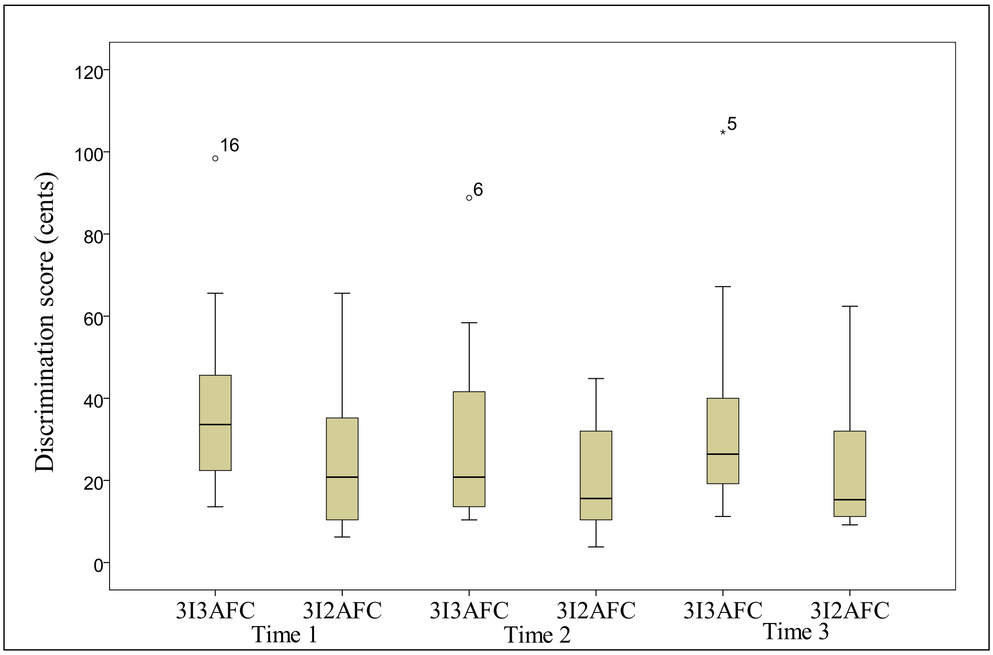


**a. Sine tone**

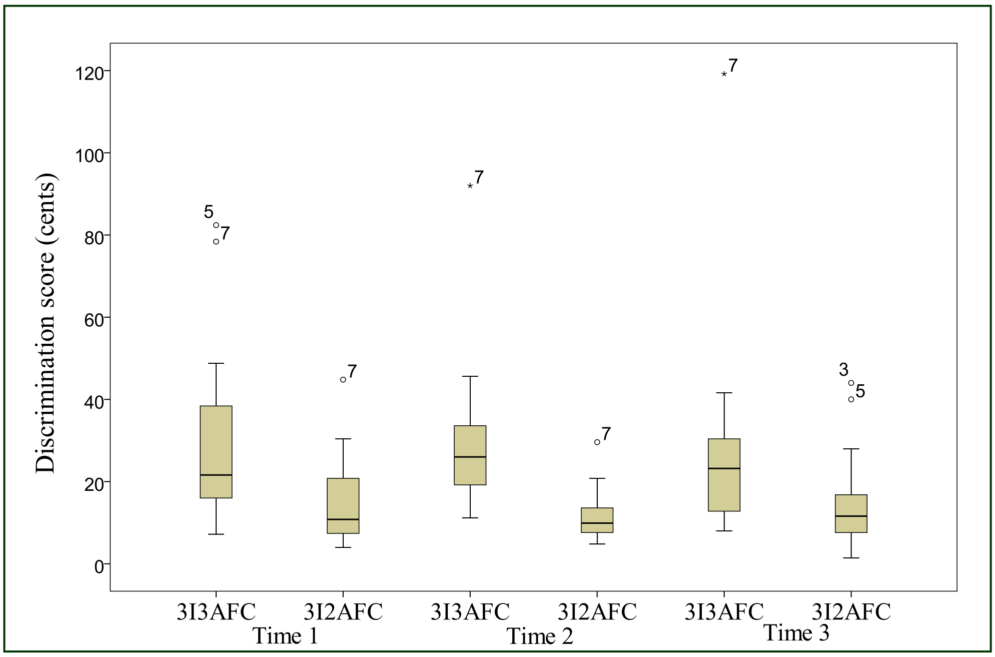


**b. Piano tone**

*Figure 3: Participants' scores from the sine tone (a) and piano tone (b) tests including the 3I3AFC procedure and the 3I2AFC procedure at time 1, time 2 and time 3. Plots show median discrimination scores and interquartile ranges. Error bars show minimum and maximum values excluding outliers.*



**a. Sine tone**



**b. Piano tone**

*Figure 4: Participants' pitch discrimination scores from the sine tone (a) and piano tone (b) tests with participants 8 and 9 removed. Plots show median discrimination scores and interquartile ranges. Error bars show minimum and maximum values excluding outliers.*

A comparison of Table 2 and Table 3 indicates that some of the mean discrimination scores and standard deviations were quite a lot larger when participants 8 and 9 were included compared to when they were removed from the data. This is most apparent for sine tone 3I3AFC time 1 scores.

*Table 2: Means and standard deviations (SD) before participants 8 and 9 were removed.*

| **Condition** | **Time 1** | | **Time 2** | | **Time 3** | |
|---|---|---|---|---|---|---|
| | **Mean (cents)** | **SD** | **Mean (cents)** | **SD** | **Mean (cents)** | **SD** |
| **Sine tone 3I3AFC** | 70.44 | 103.18 | 45.02 | 51.75 | 35.38 | 23.60 |
| **Sine tone 3I2AFC** | 27.37 | 18.05 | 20.87 | 13.02 | 33.79 | 48.78 |
| **Piano tone 3I3AFC** | 43.44 | 55.46 | 39.30 | 37.48 | 37.16 | 40.66 |
| **Piano tone 3I2AFC** | 20.29 | 22.40 | 13.38 | 8.38 | 19.23 | 17.75 |

*Table 3: Means and standard deviations (SD) after participants 8 and 9 were removed.*

| **Condition** | **Time 1** | | **Time 2** | | **Time 3** | |
|---|---|---|---|---|---|---|
| | **Mean (cents)** | **SD** | **Mean (cents)** | **SD** | **Mean (cents)** | **SD** |
| **Sine tone 3I3AFC** | 37.91 | 21.83 | 30.47 | 21.47 | 34.24 | 23.70 |
| **Sine tone 3I2AFC** | 25.21 | 17.51 | 20.17 | 13.47 | 24.39 | 18.73 |
| **Piano tone 3I3AFC** | 30.18 | 21.55 | 29.71 | 18.29 | 28.04 | 25.09 |
| **Piano tone 3I2AFC** | 15.61 | 10.46 | 11.67 | 6.31 | 15.19 | 11.57 |

Figure 4 and Table 3 suggest that participants consistently achieved lower pitch discrimination scores using the 3I2AFC procedure compared to the 3I3AFC procedure, for both the sine tone and piano tone tests (higher performance is reflected in lower pitch discrimination scores).

The Shapiro-Wilk test was carried out to investigate whether the data from the pitch tests was normally distributed. The Shapiro-Wilk test was chosen to assess if the data was normally distributed because it is said to be the most accurate test for sample sizes of less than 50 participants (Field, 2005). Table 4 shows that most of the conditions were not normally distributed ($p$<0.05).

*Table 4: Shapiro-Wilk test results for the pitch discrimination data.*

| Condition | | Degrees of Freedom (df) | Statistic | Significance |
|---|---|---|---|---|
| **3I3AFC** | Piano time 1 | 18 | .82 | .003 |
| | Piano time 2 | 18 | .76 | .000 |
| | Piano time 3 | 18 | .65 | .000 |
| | Sine time 1 | 18 | .88 | .021 |
| | Sine time 2 | 18 | .85 | .007 |
| | Sine time 3 | 18 | .83 | .004 |
| **3I2AFC** | Piano time 1 | 18 | .86 | .013 |
| | Piano time 2 | 18 | .85 | .007 |
| | Piano time 3 | 18 | .84 | .005 |
| | Sine time 1 | 18 | .90 | .066 |
| | Sine time 2 | 18 | .91 | .069 |
| | Sine time 3 | 18 | .75 | .000 |

### Main effects and interactions

Although most of the data was not normally distributed, a three-way repeated measures ANOVA was computed in order to assess the possibility of main effects and interactions. This was because there is no non-parametric equivalent of a three way repeated measures ANOVA. These results must be viewed with caution and are necessarily followed up with non parametric tests.

Mauchley's test indicated that the assumption of sphericity had been violated, $x^2$ (2) = 0.69, $p$<0.05, therefore degrees of freedom were corrected using the Greenhouse-Geiser estimate of sphericity ($\varepsilon$ = 0.76). The results of the three way repeated measures ANOVA suggested that there was not a main effect of time, ($F$ (1.53, 25.92) = 1.59, $p$>0.05, $r$ = 0.18). However, there was a significant main effect of procedure, ($F$ (1, 17) = 35.23, $p$<0.01, $r$ = 0.81) and there was a significant main effect of instrument, ($F$ (1, 17) = 8.842, $p$<0.01, $r$ = 0.55). There were no significant interactions.

In order to reduce the number of comparisons for non parametric tests, averages were calculated across conditions so that main effects could be explored via three tests. Table 5 shows that most of the averaged data was not normally distributed, so one Friedman test and two Wilcoxon signed ranks tests were used to confirm the findings of the ANOVA. A Bonferroni correction was used in order to produce more conservative $p$ values for the three comparisons. Therefore a $p$ value of 0.05/3 = 0.0167 was implemented.

*Table 5: Shapiro-Wilk test results for average scores across conditions.*

| Condition | Degrees of Freedom (df) | Statistic | Significance |
|---|---|---|---|
| **Time 1** | 18 | 0.85 | 0.009 |
| **Time 2** | 18 | 0.88 | 0.028 |
| **Time 3** | 18 | 0.86 | 0.011 |
| **3I3AFC** | 18 | 0.90 | 0.06 |
| **3I2AFC** | 18 | 0.93 | 0.18 |
| **Sine tone** | 18 | 0.95 | 0.37 |
| **Piano tone** | 18 | 0.78 | 0.001 |

### 3.1.1 Effect of time

A Friedman test was computed to check that there were no significant differences between the three test sessions using a non parametric test. The results showed that there were no statistically significant differences between the three test sessions, ($x^2$ = 4.78, *df* = 2, *p*>0.05). Both the ANOVA and the non parametric Friedman test suggested that there were no significant differences between the three test sessions.

**Correlation**

To further investigate the reliability of the two different procedures, correlations were computed between time 1, time 2 and time 3, for the two different procedures. Most of the data was not normally distributed (see Table 6) so a nonparametric Spearman's rank correlation coefficient was computed.

*Table 6: Shapiro-Wilk test results for the two procedures.*

| Condition | | Degrees of Freedom (df) | Statistic | Significance |
|---|---|---|---|---|
| **3I3AFC** | Time 1 | 18 | .90 | .049 |
| | Time 2 | 18 | .89 | .032 |
| | Time 3 | 18 | .82 | .003 |
| **3I2AFC** | Time 1 | 18 | .84 | .006 |
| | Time 2 | 18 | .90 | .057 |
| | Time 3 | 18 | .86 | .007 |

The results showed that there was a statistically significant moderate to strong positive correlation between 3I3AFC discrimination scores at time 1 and scores at time 2, *r* = .69, *p*<0.01, (one tailed). There was also a significant moderate to strong positive correlation between time 1 and time 3, *r* = 0.66, *p*<0.01, (one tailed) and between time 2 and time 3, *r* =

.68, p<0.001, (one tailed). The same tests were carried out on the 3I2AFC pitch discrimination scores for time 1, time 2 and time 3. The results showed that there was a significant positive correlation between time 1 and time 2, *r* = .70, *p*<0.001, (one tailed). Time 1 and time 3 also correlated significantly, *r* = .52, *p*<0.05, (one tailed) and time 2 and time 3 also correlated significantly, *r* = .42, *p*<0.05, (one tailed).

This shows that both procedures produced scores which correlated significantly over the three test sessions. This reaffirms the lack of a main effect of time found in the ANOVA and Friedman test. The 3I3AFC procedure showed correlations to a higher level of significance so may indicate slightly more consistent results compared to the 3I2AFC procedure.

### 3.1.2 Effect of procedure

Once the groups were averaged across the other conditions for the 3I3AFC procedure and the 3I2AFC procedure, the Shapiro-Wilk test showed the data to be normally distributed (see Table 5). A paired samples t-test was computed to find out if there was a statistically significant difference between the two different procedures. The results showed that the participants obtained significantly lower discrimination scores for the 3I2AFC procedure (Mean = 18.71 cents, SD = 9.22) compared to the 3I3AFC procedure (Mean = 31.76 cents, SD = 15.75), (*t* (17) = 5.94, *p*<0.001, *r* = 0.82). This was using a Bonferroni adjusted significance level of 0.05/3 = 0.0167.

Lower discrimination scores for the 3I2AFC procedure were as expected as participants had a higher chance of selecting the correct stimulus. According to Gescheider (1997, p147), when the proportion of correct responses p(c) is held constant at 0.707 (using the two down one up procedure), d' (which is a measure of detectability) increases as the number of observation intervals increases. Klein (2001) stated that d’ is approximately linearly related to signal strength in discrimination tasks. Therefore, values of detectability can be obtained from Gescheider (1997). This states that, if p(c) = 0.707, then d' = 1.28 for a 3AFC procedure and d' = 0.78 for a 2AFC procedure. This means that a task using a 3AFC procedure needs a higher level of signal detectability compared to a 2AFC procedure. Therefore the difference in signal detectability for this comparison was 1.64. According to this theory, if the 3I2AFC results from this experiment were multiplied by 1.64, they should not differ significantly from the 3I3AFC procedure results. If they still differed significantly then there may have been a response bias or these assumptions were not correct.

**Effect of procedure using transformed data**

Based on this assumption, the 3I2AFC results were multiplied by 1.64 and then compared against the 3I3AFC results. Table 7 shows that the data was normally distributed.

*Table 7: Shapiro-Wilk test results for the transformed 3I2AFC data and original 3I3AFC data.*

| Condition | Degrees of Freedom (df) | Statistic | Significance |
|---|---|---|---|
| **Transformed 3I2AFC** | 18 | .93 | .18 |
| **Original 3I3AFC** | 18 | .90 | .055 |

A paired samples t-test was computed to find out if there was a significant difference between the transformed 3I2AFC procedure and the original 3I3AFC procedure. The results suggested that there was not a statistically significant difference between the transformed 3I2AFC procedure (Mean = 30.68 cents, SD = 15.12) compared to the original 3I3AFC procedure (Mean = 31.76 cents, SD = 15.75), ($t$ (17) = 0.537, $p$>0.05, $r$ = 0.13). This shows that the groups were not significantly different once the 3I2AFC data had been multiplied by the correction factor. Suggesting if there was a response bias, this was not to an extent that it influenced the results.

### 3.1.3 Effect of instrument

A Wilcoxon signed ranks test was computed to find out if there was a significant difference between instruments in the pitch discrimination tests. The results suggested that participants achieved significantly lower discrimination scores in the piano tone condition (Mdn = 17.63 cents) compared to the sine tone condition (Mdn = 26 cents), ($z$ = -2.64, $p$<0.001, $r$ = 0.62), based on a Bonferroni adjusted significance criterion of 0.05/3 = 0.0167.

### 3.1.4 Comparison of pitch discrimination with a previous study

In order to examine if the data collected in the present paper was similar to data collected previously, a comparison was made between Paynter's (2010) data and the present papers data. There were some differences in the testing procedure of Paynter's (2010) study which need to be taken into account in this analysis. Paynter (2010) tested participants pitch discrimination scores in two tests. These were a high range pitch test and a low range pitch test assessed in one test session. For the purpose of the present paper, an average of the score from both the high range and the low range tests was calculated to produce one pitch discrimination score. This was compared against the time 1 3I3AFC pitch discrimination scores collected in the present paper. Paynter (2010) used an adaptive procedure with six reversals so scores were averaged over the last four reversals. This is in comparison to the present paper which used seven reversals and averaged the last five reversals. This may have had an effect on the threshold estimate. Figure 6 shows the discrimination scores from the present paper compared to Paynter's (2010) non-musician and musician data.

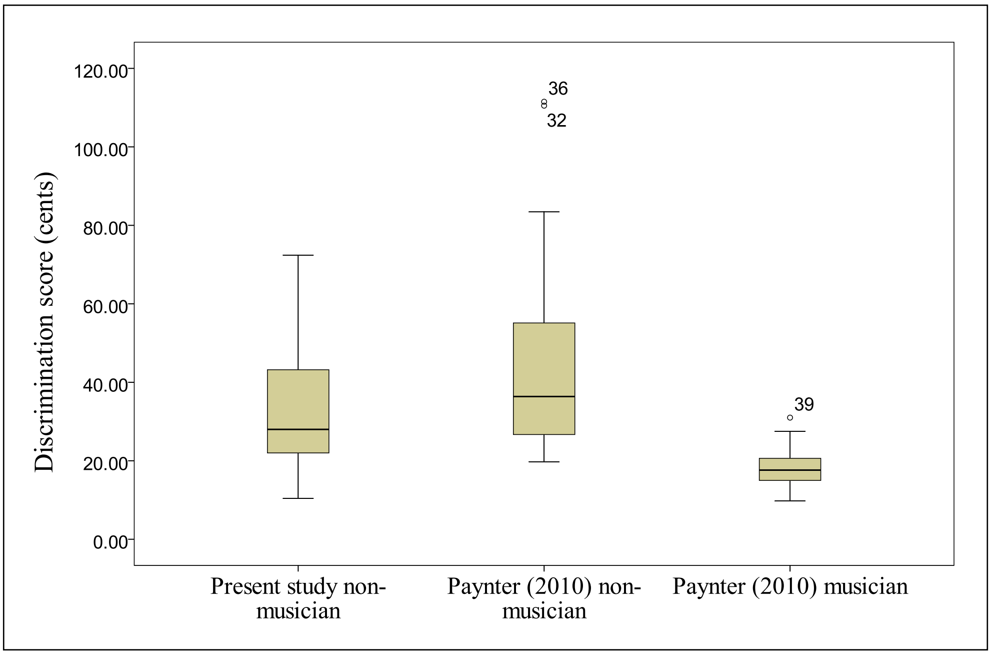


*Figure 6: Pitch discrimination scores from the present study's 3I3AFC time 1 non-musician data and Paynter's (2010) non-musician data and musician data. Plots show median discrimination scores and interquartile ranges. Error bars show minimum and maximum values excluding outliers.*

Shapiro-Wilk tests showed that some of the data was not normally distributed (see Table 12), therefore non parametric independent measures tests were computed to assess if there were any significant differences between the three groups of participants.

*Table 12: Shapiro-Wilk test results for the present paper's data and Paynter's (2010) data.*

| **Condition** | **Degrees of Freedom (df)** | **Statistic** | **Significance** |
|---|---|---|---|
| **Present papers non-musician data** | 18 | .90 | .049 |
| **Paynter's (2010) non-musician data** | 20 | .80 | .001 |
| **Paynter's (2010) musician data** | 20 | .96 | .49 |

A Kruskall-Wallis test was computed to find out if there was a statistically significant difference between the non-musician data for the present study (Mdn = 28%) compared to Paynter's (2010) non-musician (Mdn = 36.38%) and musician data (Mdn = 17.63%). The results suggested that the three groups of results were significantly different ($x^2$ (2) = 26.81, $p$<0.001).

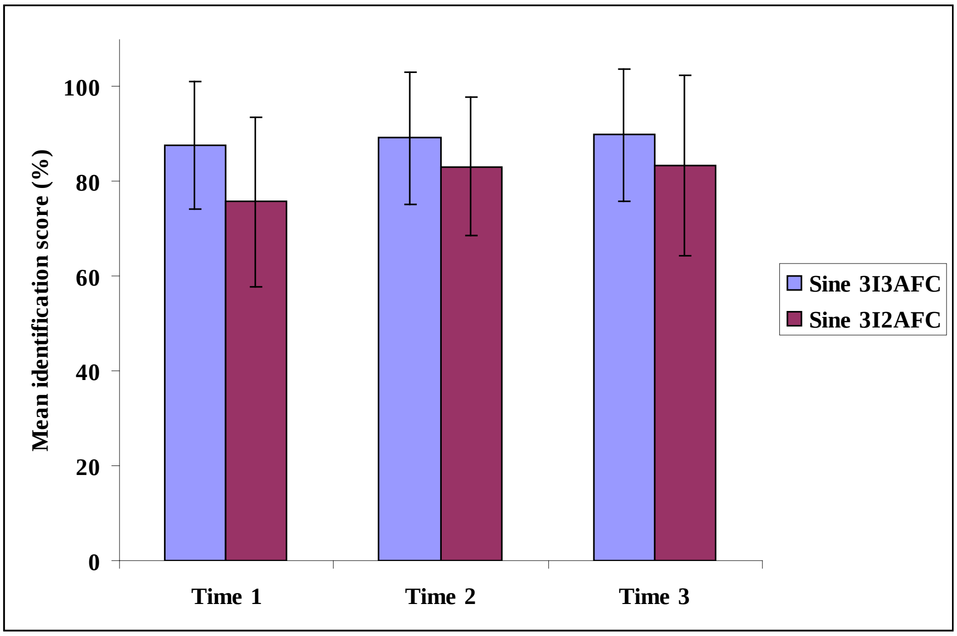
Mann Whitney U post hoc tests were computed to assess which of these groups were statistically significantly different. The results showed that the non-musician data for the present study did not differ significantly from the Paynter's (2010) non musician data, $U$ = 131, $p$>0.05, $r$ = 0.23, (two-tailed). As would be expected, the non-musician data from this study showed to be significantly different to Paynter's (2010) musician data, $U$ = 51, $p$<0.001, $r$ = 0.61, (two-tailed). This agrees with Paynters (2010) findings which showed non-musicians discrimination scores to be significantly lower compared to musicians' discrimination scores. The present analysis of Paynter's results confirmed this, $U$ = 22, $p$<0.001, $r$ = 0.76, (two-tailed).

## 3.2 Pitch identification

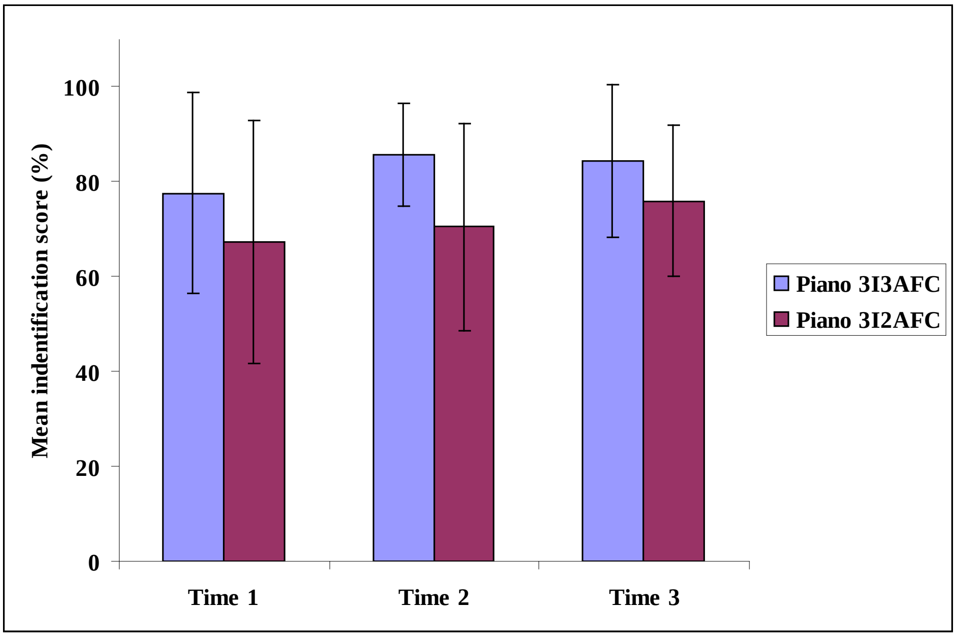
Pitch identification scores were calculated as a percentage of the number of correct pitch identifications over the last 5 reversals of the adaptive procedure. The mean pitch identification scores are shown in Table 8 and Figure 5 below. The data indicates that participants consistently achieved higher identification scores using the 3I3AFC procedure compared to the 3I2AFC procedure for both sine tone and piano tone tests.

*Table 8: Means and standard deviations (SD) after participants 8 and 9 were removed.*

| Condition | Time 1 | | Time 2 | | Time 3 | |
|---|---|---|---|---|---|---|
| | Mean (%) | SD | Mean (%) | SD | Mean (%) | SD |
| **Sine tone 3I3AFC** | 87.65 | 13.38 | 89.19 | 13.96 | 89.88 | 14.01 |
| **Sine tone 3I2AFC** | 75.72 | 17.88 | 83.21 | 14.74 | 83.34 | 19.12 |
| **Piano tone 3I3AFC** | 77.60 | 21.21 | 85.84 | 10.85 | 84.28 | 16.08 |
| **Piano tone 3I2AFC** | 67.34 | 25.63 | 70.45 | 21.85 | 76.00 | 16.05 |



**a. Sine tone**



**b. Piano tone**

*Figure 5: Mean pitch identification scores for the sine tone (a) and piano tone (b) tests using the 3I3AFC procedure and the 3I2AFC procedure. Error bars show ± 1 SD.*

Shapiro-Wilk tests were computed in order to assess whether the identification data was normally distributed (see Table 9).

*Table 9: Shapiro-Wilk test results for the pitch identification data.*

| Condition | Degrees of Freedom (df) | Statistic | Significance |
|---|---|---|---|
| **3I3AFC** Piano time 1 | 18 | .88 | .025 |
| Piano time 2 | 18 | .93 | .23 |
| Piano time 3 | 18 | .88 | .021 |
| Sine time 1 | 18 | .84 | .006 |
| Sine time 2 | 18 | .77 | .001 |
| Sine time 3 | 18 | .75 | .00 |
| **3I2AFC** Piano time 1 | 18 | .93 | .16 |
| Piano time 2 | 18 | .94 | .29 |
| Piano time 3 | 18 | .95 | .41 |
| Sine time 1 | 18 | .96 | .51 |
| Sine time 2 | 18 | .93 | .16 |
| Sine time 3 | 18 | .83 | .005 |

**Main effects and interactions**

Although most of the data was not normally distributed, a three-way repeated measures ANOVA was computed in order to assess the possibility of main effects and interactions. This was carried out because there is no non parametric equivalent to the three way repeated measures ANOVA. These results need to be viewed with caution and were followed up with non parametric tests.

Mauchley’s test indicated that the assumption of sphericity could be assumed, $x^2(2) = 0.931$, p>0.05. The results of the three-way repeated measures ANOVA suggested that there was a main effect of time, ($F$ (2,34) = 3.93, $p$<0.05, $r$ = 0.37). There was also statistically significant main effect of procedure, ($F$ (1,17) = 29.35, $p$<0.001, $r$ =0.78 ) and there was a significant main effect of instrument, ($F$ (1,17) = 28.46, $p$<0.01, $r$ = 0.78). No other comparisons were significant.

In order to reduce the number of comparisons for non parametric tests, averages were calculated across conditions so that main effects could be explored via three tests. The averaged scores showed to be normally distributed (see Table 10) so a one-way repeated measures ANOVA and two t-tests were used to confirm the findings of the three-way ANOVA. A Bonferroni adjusted $p$ value of 0.05/3 = 0.0167 was implemented.

*Table 10: Shapiro-Wilk test results for the average identification scores.*

| Condition | Degrees of Freedom (df) | Statistic | Significance |
|---|---|---|---|
| **Time 1** | 18 | .96 | .57 |
| **Time 2** | 18 | .94 | .28 |
| **Time 3** | 18 | .95 | .45 |
| **3I3AFC** | 18 | .93 | .18 |
| **3I2AFC** | 18 | .94 | .24 |
| **Sine** | 18 | .95 | .37 |
| **Piano** | 18 | .93 | .19 |

### 3.2.1 Effect of time

A one-way repeated measures ANOVA was computed to assess if there was a significant difference between time 1, time 2 and time 3 for the averaged scores. Mauchley's test indicated that the assumption of sphericity could be assumed, $x^2(2) = 0.931$, $p>0.05$. The results of the one way repeated measures ANOVA showed that there was not a main effect of time, ($F(2,34) = 3.93$, $p>0.0167$, $r = 0.37$). This shows that once the Bonferroni correction has been made for the number of comparisons, the three groups were not significantly different.

**Correlation**

The results suggest that there were no significant differences between time 1, time 2 and time 3. Shapiro-Wilk tests showed that some of the data was not normally distributed for the 3I3AFC procedure (see Table 11). Therefore, a non parametric spearman's rank correlation coefficient was computed in order to identify if participants 3I3AFC identification scores correlated between time 1, time 2 and time 3.

*Table 11: Shapiro-Wilk test results for the two procedures.*

| Condition | Degrees of Freedom (df) | Statistic | Significance |
|---|---|---|---|
| **3I3AFC** Time 1 | 18 | .93 | .18 |
| Time 2 | 18 | .88 | .024 |
| Time 3 | 18 | .86 | .014 |
| **3I2AFC** Time 1 | 18 | .92 | .15 |
| Time 2 | 18 | .94 | .24 |
| Time 3 | 18 | .97 | .87 |

The results suggested that there was a statistically significant positive correlation between 3I3AFC identification scores at time 1 and scores at time 2, $r = .87$, $p<0.001$, (one tailed).

There was also a significant correlation between time 1 and time 3, *r* = .92, *p*<0.001, (one tailed). Time 2 and time 3 were also significantly correlated, *r* = .71, *p*<0.01, (one tailed).
The data for the 3I2AFC procedure was normally distributed (see Table 11) so a Pearson's r correlation was carried out on the 3I2AFC identification scores for time 1, time 2 and time 3. The results showed that there was a significant positive correlation between time 1 and time 2, *r* = .75, *p*<0.001, (one tailed). However, there was not a correlation between time 1 and time 3, *r* = .38, *p*>0.05, (one tailed). Also, time 2 and time 3 did not show a significant correlation, *r* = .38, *p*>0.05, (one tailed). This suggests that the 3I3AFC procedure showed more consistent results and therefore more test-retest reliability compared to the 3I2AFC procedure.

### 3.2.2 Effect of procedure

Once the groups were averaged across the other conditions for the 3I3AFC procedure and the 3I2AFC procedure, the Shapiro-Wilk test showed the data to be normally distributed (see Table 10). A paired samples t-test was computed to find out if there was a significant difference between the two different procedures. The results showed that the participants obtained significantly lower identification scores using the 3I2AFC procedure (Mean = 76.01 %, SD = 12.93) compared to the 3I3AFC procedure (Mean = 85.74 %, SD = 11.63), (*t* (17) = 5.42, *p*<0.001, *r* = 0.80), using a Bonferroni adjusted significance level of *p*=0.05/3.

### 3.2.3 Effect of instrument

A paired samples t-test was also computed to find out if there was a statistically significant difference between the sine tone and piano tone data, regardless of the other conditions. The results showed that the participants obtained significantly higher identification scores in the sine tone condition (Mean = 84.83%, SD = 10.45) compared to the piano tone condition (Mean = 76.92%, SD = 13.56), (*t* (17) = 5.33, *p*<0.001, *r* = 0.79).

**Number of correct identifications above chance level**

Further analysis of the data was carried out in order to assess the proportion of correct identifications in the last five reversals for each test trial. Subsequently, a calculation was made to identify if the proportion of correct responses for each trial occurred at a level above chance based on a probability level of *p*<0.05. The data showed that for the sine 3I3AFC procedure, 11/18 (time 1), 15/18 (time 2) and 17/18 (time 3) participants made correct identifications at a level above chance (*p*<0.05). For the sine 3I2AFC procedure, 9/18 (time

1), 12/18 (time 2) and 13/18 participants made correct identifications at a level above chance (*p*<0.05).

The data showed that for the piano 3I3AFC procedure, 9/18 (time 1), 13/18 (time 2) and 12/18 (time 3) participants made correct identifications at a level above chance (*p*<0.05). For the piano 3I2AFC procedure 8/18 (time 1), 8/18 (time 2) and 9/18 (time 3) participants made correct identifications at a level above chance (*p*<0.05).

### 3.2.4 Comparison of pitch identification with a previous study

The same comparison with Paynter (2010) was made for the identification scores. Differences in the methodology need to be taken into account for the identification scores in the same way as the discrimination scores. Paynter's (2010) high range and low range scores were averaged to produce one pitch identification score for the purpose of this analysis. Another difference was that Paynter (2010) implemented an adaptive procedure with six reversals and identification scores were calculated as a percentage over the last four reversals. The present paper used seven reversals and calculated the percentage correct score over the last five reversals. This means that comparisons must be viewed with caution as participants were not exposed to the same testing procedure. Figure 7 shows mean pitch identification scores for the present papers 3I3AFC time 1 data compared to Paynter's (2010) non-musician and musician data.

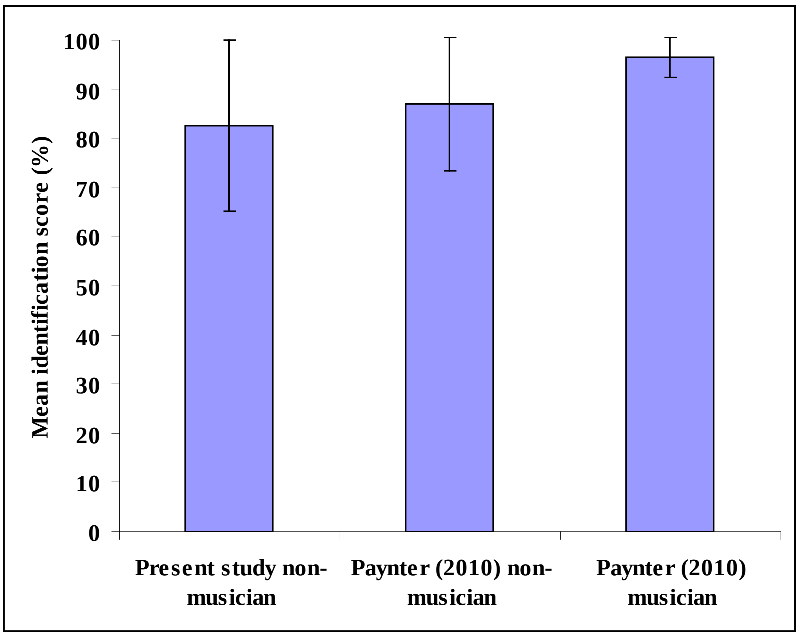


*Figure 7: Mean identification scores for the three groups. Error bars show +/- 1 SD.*

Shapiro-Wilk tests showed two of the groups were normally distributed (see Table 13) so non parametric independent measures tests were computed.

*Table13: Shapiro-Wilk test results for the pitch identification scores.*

| **Condition** | **Degrees of Freedom (df)** | **Statistic** | **Significance** |
|---|---|---|---|
| **Present papers non-musician data** | 18 | .92 | .12 |
| **Paynter's (2010) non-musician data** | 20 | .80 | .001 |
| **Paynter's (2010) musician data** | 20 | .88 | .02 |

A Kruskall-Wallis test was computed to find out if there was a significant difference between the non-musician data from the present study (Mdn = 84.73 %) compared to Paynter's (2010) non-musician (Mdn = 90.08 %) and musician data (Mdn = 96.88 %). The results showed that the three groups of results were significantly different ($x^2$ (2) = 16.51, $p$<0.001).

Mann Whitney U post hoc tests were computed to assess which of these groups were statistically significantly different. The results showed that the non-musician data for the present study did not differ significantly from the Paynter's (2010) non musician data, $U$ = 149.50, $p$>0.05, $r$ = 0.14, one-tailed). As would be expected, the non-musician data from this study showed to be significantly different to Paynter's (2010) musician data, $U$ = 62, $p$<0.001, $r$ = 0.56, one-tailed. This agrees with Paynters (2010) findings which showed non-musicians' discrimination scores to be significantly lower compared to musicians' discrimination scores. The present analysis of Paynter's results confirmed this, ($U$ = 73.50, $p$<0.001, $r$ = 0.56, one-tailed).

## 3.3 Rhythm discrimination

Rhythm discrimination scores were analysed to assess whether participants achieved different results when tested with the two different procedures and during the three repeated test sessions. Table 14 and Figure 8 suggest that participants achieved higher discrimination scores in the 3I3AFC condition compared to the 3I2AFC condition.

*Table 14: Means and standard deviations (SD) for rhythm discrimination data.*

| Condition | Time 1 | | Time 2 | | Time 3 | |
|---|---|---|---|---|---|---|
| | Mean (secs) | SD | Mean (secs) | SD | Mean (secs) | SD |
| **Rhythm 3I3AFC** | 0.079 | 0.061 | 0.066 | 0.053 | 0.059 | 0.037 |
| **Rhythm 3I2AFC** | 0.050 | 0.077 | 0.028 | 0.013 | 0.026 | 0.013 |

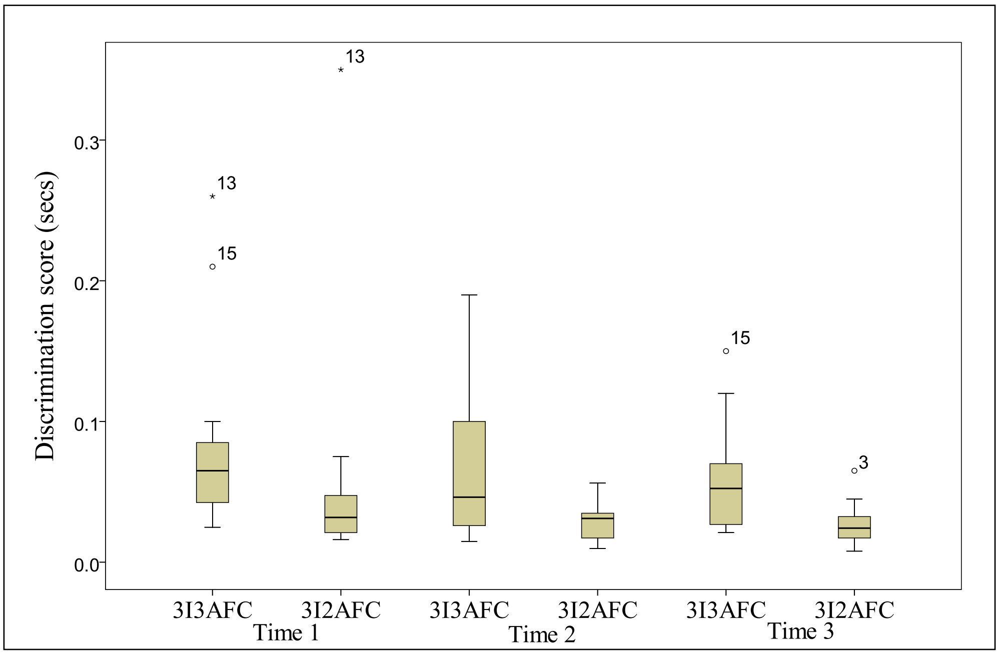


*Figure 8: Box plots showing participants scores from the rhythm tests comparing the 3I3AFC procedure and the 3I2AFC procedure at time 1, time 2 and time 3. Plots show median discrimination scores and interquartile range. Error bars show minimum and maximum values excluding outliers.*

The Shapiro-Wilk tests showed that most of the data was not normally distributed (see Table 15).

*Table 15: Shapiro-Wilk test results for the rhythm discrimination data.*

| Condition | | Degrees of Freedom (df) | Statistic | Significance |
|---|---|---|---|---|
| **3I3AFC** | Time 1 | 18 | .73 | .00 |
| | Time 2 | 18 | .83 | .004 |
| | Time 3 | 18 | .88 | .025 |
| **3I2AFC** | Time 1 | 18 | .43 | .00 |
| | Time 2 | 18 | .92 | .15 |
| | Time 3 | 18 | .88 | .022 |

**Main effects and interactions**

Although the data was not normally distributed, a two-way repeated measures ANOVA was computed. This was because there is no non parametric equivalent of a two-way repeated measures ANOVA.

Mauchley's test indicated that the assumption of sphericity had been violated, $x^2(2) = 0.24$, $p<0.05$, therefore degrees of freedom were corrected using the Greenhouse-Geiser estimate of sphericity ($\varepsilon$ =0.57). The results of the two way repeated measures ANOVA showed that there was not a main effect of time, ($F(1.14, 19.37) = 2.12$, $p>0.05$, $r = 0.28$). However, there was a significant main effect of procedure, ($F$ (1,17) = 23.41, $p<0.001$, $r = 0.97$). No other comparisons were significant.

In order to reduce the number of comparisons for non parametric tests, averages were calculated across conditions so that main effects could be explored via two tests. Shapiro-Wilk tests suggested that the data was not normally distributed, therefore a Friedman test and a Wilcoxon signed ranks test were computed to confirm the findings of the two-way ANOVA. The significance criterion was adjusted using the Bonferroni correction, therefore a $p$ value of 0.05/2 = 0.025 was used.

Shapiro-Wilk tests showed that the averaged rhythm discrimination scores were not normally distributed (see Table 16).

*Table 16: Shapiro-Wilk test results for the averaged rhythm discrimination scores.*

| Condition | Degrees of Freedom (df) | Statistic | Significance |
|---|---|---|---|
| **Time 1** | 18 | .57 | .00 |
| **Time 2** | 18 | .89 | .040 |
| **Time 3** | 18 | .89 | .039 |
| **3I3AFC** | 18 | .86 | .014 |
| **3I2AFC** | 18 | .60 | .00 |

### 3.3.1 Effect of time

A Friedman test was computed to check that there were no significant differences between the three test sessions using a non parametric test. The results suggested that there were no statistically significant differences between the three test sessions, ($x^2$ = 5.44, $df$ = 2, $p>0.05$).

### Correlation

The results suggest that there were no significant differences between time 1, time 2 and time 3. Therefore, a non parametric Spearman's rank correlation coefficient was computed in order to identify if participants rhythm discrimination scores correlated between time 1, time 2 and time 3. The results showed that there was a significant positive correlation between 3I3AFC rhythm discrimination scores at time 1 and scores at time 2, $r$ = .69, $p$<0.001, (one tailed). There was also a significant correlation between time 1 and time 3, $r$ = .65, $p$<0.01, (one tailed). Time 2 and time 3 were also significantly correlated, $r$ = .76, $p$<0.001, (one tailed).

The same tests were carried out on the 3I2AFC identification scores for time 1, time 2 and time 3. The results showed that there was not a significant positive correlation between time 1 and time 2, $r$ = .15, $p$>0.05, one tailed. In addition, there was not a significant correlation between time 2 and time 3, $r$ = .29, $p$>0.05, (one tailed). However, time 1 and time 3 showed a significant correlation, $r$ = .41, $p$<0.05, (one tailed). This suggests that the 3I3AFC procedure may show more test-retest reliability, because the 3I2AFC procedure showed some non significant correlations between test sessions.

### 3.3.2 Effect of procedure

A non parametric Wilcoxon signed rank test confirmed the results from the ANOVA. The results showed that participants achieved significantly lower rhythm discrimination scores in the 3I2AFC condition (Mdn = 0.028 secs) compared to the 3I3AFC condition (Mdn = 0.057 secs), ($z$ = -3.72, $p$<0.001, $r$ = 0.88).

### Effect of procedure with transformed 3I2AFC data

In order to correct for the difference in the level of detectability for the 3I2AFC procedure, the 3I2AFC data was multiplied by 1.64 (this was explained in more detail in section 3.1). Shapiro-Wilk tests showed that the data was not normally distributed (see Table 17), so a Wilcoxon signed ranks test was computed to find out if there was a significant difference between the transformed 3I2AFC procedure and the original 3I3AFC procedure.

*Table 17: Shapiro-Wilk test results for the transformed 3I2AFC rhythm discrimination scores and original 3I3AFC scores.*

| Condition | Degrees of Freedom (df) | Statistic | Significance |
|---|---|---|---|
| **Transformed 3I2AFC** | 18 | .60 | .00 |
| **Original 3I3AFC** | 18 | .86 | .014 |

The results of the Wilcoxon signed ranks test suggested that there was not a significant difference between the transformed 3I2AFC procedure (Mdn = 0.045) compared to the original 3I3AFC procedure (Mdn = 0.057), ($z$ = -1.42, $p$>0.05, $r$ = 0.33). This shows that the groups were not significantly different once the 3I2AFC data had been multiplied by the 1.64 correction factor.

### 3.3.3 Comparison of rhythm discrimination with a previous study

In order to examine if the rhythm discrimination data collected in the present paper was similar to data collected previously, a comparison was made between Paynter's (2010) data and the present papers data. A comparison was made between the present papers time 1 3I3AFC data and Paynter's (2010) non-musician and musician data. Testing for the three different groups was similar however, as discussed earlier; Paynter (2010) used an adaptive procedure with six reversals so scores were averaged over the last four reversals. This is in comparison to the present research project which used seven reversals and averaged the last five reversals. This may have had an effect on the threshold estimate. The data is shown below in Figure 9. There appears to be a lot less variability in Paynter's (2010) musician data compared to Paynter's (2010) non-musician data and the present papers non-musician data.

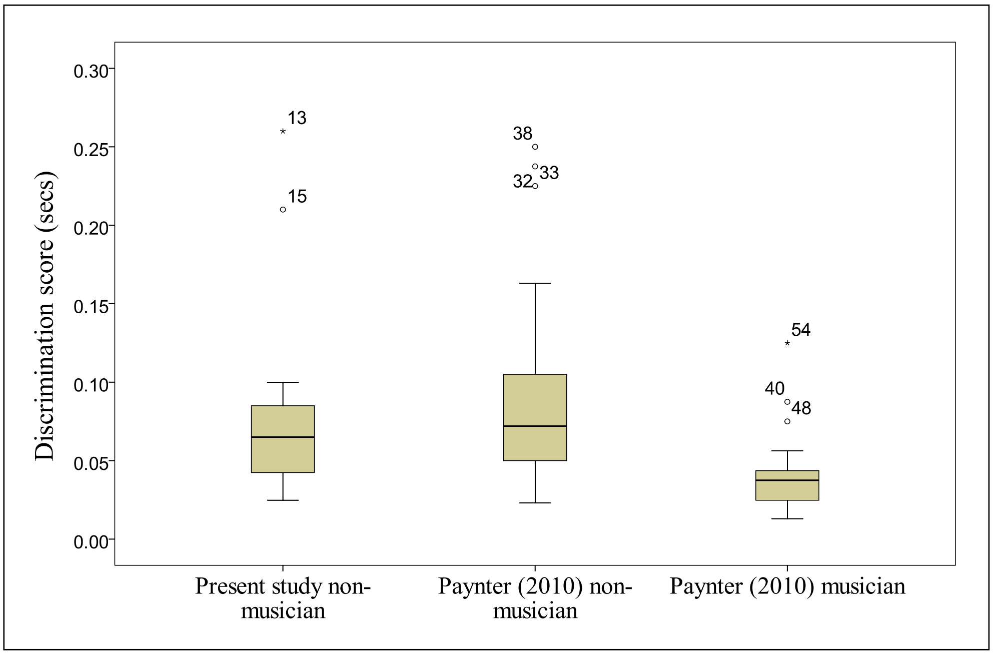


*Figure 9: The present papers 3I3AFC rhythm discrimination data at time 1 and Paynter's (2010) non-musician data and musician data. Plots show median discrimination scores and interquartile range. Error bars show minimum and maximum values excluding outliers.*

Shapiro-Wilk tests suggested that none of the data was normally distributed (see Table 18), therefore non parametric independent measures tests were computed.

*Table 18: Shapiro-Wilk test results for the three different groups.*

| **Condition** | **Degrees of Freedom (df)** | **Statistic** | **Significance** |
|---|---|---|---|
| **Present papers non-musician data** | 18 | .73 | .00 |
| **Paynter's (2010) non-musician data** | 20 | .81 | .001 |
| **Paynter's (2010) musician data** | 20 | .79 | .001 |

A Kruskall-wallis test was computed to find out if there was a significant difference between the present study's non-musician (Mdn = 0.065 secs) data compared to Paynter's (2010) non-musician (Mdn = 0.072 secs) and musician data (Mdn = 0.038 secs). The results showed that the three groups of results were significantly different ($x^2$ (2, 58) = 12.11, $p$<0.01).

Mann Whitney U post hoc tests were computed to assess which of these groups were significantly different. The results showed that the non-musician data for the present study did not differ significantly from Paynter's (2010) non musician data, $U$ = 159.5, $p$>0.01, $r$ = 0.59, (two-tailed). As would be expected, the non-musician data from this study was shown to be significantly different to Paynter’s (2010) musician data, $U = 89.5$, $p<0.01$, $r = 0.59$, (two-tailed). This agrees with Paynters (2010) findings which showed non-musicians discrimination scores to be significantly lower compared to musicians’ discrimination scores. The present analysis of Paynter’s results confirmed this, $U = 81$, $p<0.01$, $r = 0.72$, (two-tailed).

# 4 Discussion

The aim of this research project was to assess the test-retest reliability of the SOECIC MTB and to investigate whether two different response methods influenced the test-retest reliability of the tests. Participants were required to complete pitch discrimination, pitch identification and rhythm discrimination tests (using a 3I3AFC procedure and a 3I2AFC procedure). This was repeated on three separate occasions. The analysis of the data aimed to assess whether there was a statistically significant difference between participants' scores over the three repeated sessions. In addition, analyses were performed to examine differences between the data when participants were using the 3I3AFC procedure compared to the 3I2AFC procedure. The participants were required to complete pitch tests using sine tone stimuli and piano tone stimuli, therefore an effect of instrument was also investigated. It is important to note that the lower the discrimination score, the better the performance, whereas the opposite was true for the identification scores.

## 4.1 Pitch discrimination

The results of the pitch discrimination tests showed that there was not a main effect of time (test session). Further investigation suggested that there was a moderate to strong positive correlation between the three test sessions for both the 3I3AFC and 3I2AFC procedures. These results suggest that the measurement of participants pitch discrimination ability was consistent over the three test sessions, thus the pitch discrimination tests showed good test-retest reliability. However, the correlations were not as highly statistically significant for the 3I2AFC procedure compared to the 3I3AFC procedure suggesting that the 3I3AFC procedure may show slightly more consistent results.

The data showed that on average over the three test sessions, participants obtained mean discrimination scores of 29.31 cents (piano tone 3I3AFC), 14.16 cents (piano 3I2AFC), 34.21 cents (sine tone 3I3AFC) and 23.26 cents (sine tone 3I2AFC). These all equate to less than one semitone (100 cents) which is consistent with results found in previous literature examining normal hearing participants (for example, Gfeller et al., 2002, cited in section 1.2.2). Many of the studies examining music perception in cochlear implant users did not test at levels below one semitone.

There has been some research indicating that a number of CI users are capable of discriminating differences of less than one semitone (Gfeller et al., 2002). Also, Gfeller et al. (2007) showed that some CI users who were using CI implants with shorter electrode arrays and had the use of residual acoustical hearing, performed better in pitch perception tasks than CI users with electrical hearing only (discussed previously in section 1.2.2). The development of hybrid and bi-modal CIs which make use of residual acoustical hearing as well as electrical stimulation mean that it is likely that CI users will be able to discriminate smaller pitch differences. The fact that the SOECIC MTB has the scope to measure much smaller thresholds means that it will also be suitable in the future as further advances in CI technology are made.

The statistical analysis showed that there was a significant main effect of procedure. The findings indicated that participants achieved lower pitch discrimination scores for the 3I2AFC procedure compared to the 3I3AFC procedure. This is what was expected as participants would have had a higher chance of making the correct response when presented with the 3I2AFC procedure. As explained in section 3.1.2, according to Geschieder (1997, p.147), when the proportion of correct responses p(c) is held constant at 0.707 (using the 2-down 1-up procedure), d' (which is a measure of detectability) increases as the number of observation intervals increases. Values of d’ were obtained from Geschieder (1997) to calculate a correction factor. Klein (2001) explained how d’ is approximately linearly related to signal strength in discrimination tasks, assuming there is no response bias. This meant that the d’ for a 2AFC task and d’ for a 3AFC task could be taken and used to calculate what the difference in threshold was likely to be using the ratio of d’ values. This produced a d’ ratio of 1.64, therefore the 3I2AFC data was multiplied by 1.64.

The results showed that once the 3I2AFC data had been transformed by the correction factor, there was not a statistically significant difference between the groups. Klein (2001) suggested that if this correction factor did not make the two groups similar, then there may have been other factors affecting the experiment such as a response bias. It is not possible to make the assumption that there was not a response bias in this data; however, if there was a response bias, it appears that it may not have been to the extent that it influenced the outcome of the statistical analysis. The existence of a response bias in this data was not investigated due to time constraints but it would be an interesting factor to examine in future analysis of the data.

The analyses showed that there was a statistically significant main effect of instrument. The participants achieved significantly lower pitch discrimination scores in the piano tone tests compared to the sine tone tests. This could be explained by the fact that there would have been more cues available for pitch discrimination when participants were presented with piano tones compared to sine tones. A previous study examining pitch perception found that participants achieved better scores when discriminating between small changes in complex tones compared to pure tones (Zeitlin, 1964). The author suggested that the presence of overtones in the complex tone stimuli aided in the participants pitch discrimination. In the present study, the overtones which would have occurred after the initial onset of the piano tones may have acted as an extra cue which was not present for the sine tones. This implies that higher levels of musical experience could improve pitch perception ability, as the musically trained may become more able to make use of these extra cues such as overtones. The present study confirmed this through a comparison between the present study's non-musician data and a previous study's data testing musicians (Paynter, 2010). The findings indicated that the present study's non-musician pitch discrimination scores were significantly higher than Paynter's (2010) musician pitch discrimination scores.

In addition, no statistically significant difference was observed between the present study's non-musician pitch discrimination scores and Paynter's (2010) non-musician pitch discrimination scores. This suggests that the SOECIC MTB has the capability to measure pitch discrimination in two different groups of normal hearing non-musicians and produce similar results.

There were a number of differences in the methodology of Paynter's (2010) experiment so this comparison does need to be viewed with caution. However, the fact that they did not differ suggests that these differences may not have had a significant impact on the outcome of the tests. Paynter's (2010) experiment measured thresholds with an adaptive procedure implementing six reversals (excluding the first two). This is compared to the present study which measured thresholds implementing seven reversals (excluding the first two). The number of reversals could potentially have influenced the threshold measurements. If it did influence the results, it may not have been to an extent that it was shown statistically in this experiment.

Furthermore, Paynter (2010) measured high and low pitch ranges separately whereas the present study took one full range measurement. In order to compare the two sets of data, an average was taken of Paynter's (2010) high and low range pitch discrimination data to produce one pitch discrimination score. The fact that the two sets of data were treated differently means that it is not completely fair to compare them directly. However, the two sets of non-musician data did not differ statistically which suggests that this may not have had a significant effect on the scores used for analysis. It is possible however, that the differences between the present study's non-musicians and Paynter's (2010) musician's could have been a result of the differences in methodology rather than musicianship status. This seems unlikely as Paynter (2010) found musicians to perform better than non-musicians previously.

## 4.2 Pitch identification

The results of the pitch identification tests showed that there was not a main effect of time (test session). The participant's scores did not differ significantly between the three test sessions. Further analysis suggested that there were moderate to strong positive correlations between the three test sessions for the 3I3AFC procedure. However, for the 3I2AFC procedure, there was a statistically significant correlation between time 1 and time 2 but not between time 1 and time 3 or between time 2 and time 3. This suggests that the 3I3AFC procedure showed more consistent results and therefore more test-retest reliability.

The tests for the effect of procedure suggested that there was a statistically significant main effect of procedure. The participants achieved statistically higher identification scores for the 3I3AFC procedure compared to the 3I2AFC procedure. This may have been because they reached lower discrimination thresholds in the 3I2AFC condition so found it more difficult to identify the direction of pitch change.

The results also suggested that there was a statistically significant main effect of instrument. The participants achieved higher discrimination scores in the sine tone condition compared to the piano tone condition. As discussed earlier regarding the lower identification scores for the 3I2AFC condition, this may have been because they reached lower discrimination thresholds in the piano tone condition so found it more difficult to identify the direction of the pitch change at threshold. Another explanation could be that perhaps there was more ambiguity in the piano tones compared to the sine tones. This is because there are harmonic and temporal fluctuations in the spectrum of piano tones which could cause confusion; these are not found in sine tones.

The identification scores were calculated as the percentage of correct pitch identifications in the last five reversals. A problem measuring pitch identification in this way is that the percentage correct score is calculated depending on the individual's adaptive procedure. For example, one participant's adaptive procedure may move straight down to threshold and then move up and down with a relatively small number of trials. However, another participant may move up and down the difficulty levels, requiring a lot more trials to obtain the seven reversals. This means that some identification scores were calculated from four trials. Thus if a participant got 4/4 correct identifications, they would receive a score of 100%. The problem with calculating the score from four trials is that participants could obtain a 100% correct score relatively easily by chance ($p$>0.05). This brings into question how informative the identification scores were. The only way to reduce the likelihood of participants making correct identifications by chance, is to implement adaptive procedures with more reversals, so that identification scores are never calculated using less than five trials (gives $p$<0.05). The problem with more reversals is that the test would take longer to administer, which could introduce factors such as participant fatigue.

In order to make sense of the identification scores obtained in the present study, further analysis involved calculating the proportion of correct identifications in the last five reversals, that occurred at a level above chance ($p$<0.05). The data showed that for the sine 3I3AFC procedure, 11/18 (time 1), 15/18 (time 2) and 17/18 (time 3) participants made correct identifications at a level above chance ($p$<0.05). For the sine 3I2AFC procedure, 9/18 (time 1), 12/18 (time 2) and 13/18 (time 3) participants made correct identifications at a level above chance ($p$<0.05). It appears that the number of participants identifying the direction of pitch change at a level above chance increased over the test sessions, indicating a possible practice effect. Also, more participants were identifying at a level above chance when using the 3I3AFC method compared to the 3I2AFC method. This may be because participants reached lower thresholds during the 3I2AFC method, and therefore found it more difficult to identify the pitch direction. This is confirmed by the fact that statistical tests showed that participants obtained significantly higher identification scores in the 3I2AFC procedure compared to the 3I3AFC procedure.

The data showed that for the piano 3I3AFC procedure, 9/18 (time 1), 13/18 (time 2) and 12/18 (time 3) participants made correct identifications at a level above chance ($p$<0.05). For the piano 3I2AFC procedure 8/18 (time 1), 8/18 (time 2) and 9/18 (time 3) participants made correct identifications at a level above chance ($p$<0.05). Participants appeared to make fewer correct identifications for the piano tones compared to the sine tones. This may be due to the fact that participants reached a significantly smaller threshold in the piano discrimination task compared to the sine tone discrimination task, therefore found it more difficult to identify the pitch change direction. Participants reported that they could tell which one was different but were unable to identify if it was higher or lower. As discussed previously, this could also be due to the fact that there may be less ambiguity in sine tones because there are no harmonic or temporal fluctuations in the spectrum of sine tones.

In a similar way to the pitch discrimination scores, a comparison was made between the data obtained during the present research project and a research project carried out previously (Paynter, 2010). The results suggested that there was not a significant difference between the present study's non-musician pitch identification scores and Paynter's (2010) non-musician pitch identification scores. The comparison also showed the present study's non-musician identification scores were significantly lower than Paynter's (2010) musician identification scores. This was what would be expected as the present research project selected non-musicians using the same criteria as Paynter (2010).

## 4.3 Rhythm discrimination

The results of the rhythm discrimination testing showed that there was not a main effect of time. Further analysis of the results suggested that there were moderate to strong correlations between the three test sessions for the 3I3AFC procedure. However, for the 3I2AFC procedure, there was not a statistically significant correlation between time 1 and time 2 and between time 2 and time 3, but there was a statistically significant correlation between time 1 and time 3. This was consistent with the findings for the pitch discrimination and identification tests. Thus, the 3I3AFC procedure appeared to produce more consistent results compared to the 3I2AFC procedure.

Participants obtained mean rhythm discrimination scores averaged over the three test sessions of 0.068 seconds for the 3I3AFC procedure and 0.035 seconds for the 3I2AFC procedure. This equates to 68 milliseconds and 35 milliseconds respectively, which is higher than Gelfand's (1998) gap detection estimate of 2 to 3 milliseconds (stated previously in section 1.2.1). However, gap detection is a very different task to the rhythm test presented in the SOECIC MTB. It is possible that the two tasks require different skills and an ability to perform well at one does not necessarily mean someone will perform well at the other. The gap detection task requires participants only to identify a gap between two stimuli. This means that participants do not necessarily need to be able to feel a rhythm. This is in comparison to the rhythm task in the SOECIC MTB which requires participants to identify a rhythm that is different. A task such as gap detection which does not require any musical ability could be more universally useful. However, a gap detection task does not possess much ecological validity, therefore may not be a particularly helpful indicator of CI users' abilities. The rhythm test in the SOECIC MTB is based on simple musical intervals which is something CI users will encounter on a day to day basis. It could be argued that a test like this will provide more useful information regarding how well a CI user is coping by presenting them with sounds which are more likely to occur in everyday life.

In addition, the present study's measurements were taken in a sound field in non anechoic conditions, which means participants were unlikely to reach the 2 to 3 millisecond thresholds similar to gap detection tasks. The smallest rhythm discrimination threshold obtained in the present study was 0.021 seconds (21 milliseconds). The SOECIC MTB is designed for CI users, therefore the tests must be presented in a sound field.

Several of the participants reported that they found the rhythm discrimination test very difficult. It is possible that a combination of the participant finding the task difficult and the repeated exposure to the same tests could have resulted in participants not performing as well as they could have done. Previous research into human behaviour suggests that when participants are faced with a difficult task they may lose motivation and 'give-up', especially if they believe they are not succeeding at the task (Brehm and Self, 1989). The fact that the participants were not given any feedback could have resulted in this. However, this is unlikely as there were no statistically significant differences between the three test sessions. This shows that participants' performance did not appear to deteriorate throughout the test sessions, which implies that a loss in motivation was not a problem during the present experiment.

The statistical analysis suggested that there was a main effect of procedure. The participants obtained significantly lower discrimination scores for the 3I2AFC procedure compared to the 3I3AFC procedure. This is what was expected as participants would have had a higher chance of making the correct response when presented with the 3I2AFC procedure. In a similar way to the pitch discrimination tests, this was further investigated by transforming the data by a correction factor of 1.64. According to this theory the correction should make the 2AFC data comparable to the 3AFC data. The results showed that once the 3I2AFC data had been transformed by the correction factor, there was not a statistically significant difference between the groups.

In a similar way to the pitch discrimination and identification scores, a comparison was made between the data obtained during the present research project and a research project carried out previously (Paynter, 2010). The results suggested that there was not a significant difference between the present study's non-musician rhythm discrimination scores and Paynter's (2010) non-musician rhythm discrimination scores. The comparison also showed the present study's non-musician data to be significantly different from Paynter's (2010) musician data.

## 4.4 General discussion

During the testing, one participant did not continue with the experiment because it was clear that after one session, they were unable to make correct responses at the easiest level of the test. The participant was given the opportunity to complete extra practice sessions to ensure they understood the instructions, however the participant still did not improve. It was decided that this participant would not be used for the experiment as it is not possible to examine the test-retest reliability of the SOECIC MTB on someone who is not capable of making correct responses at the least difficult level.

It was clear that at the beginning of the analysis, there were some outliers in the data. Before any further analysis, the decision was made to remove two of the participants from the data (participants 8 and 9). The rationale for the removal of these participants was based on the variability in their performance in the different conditions, throughout the three test sessions. Previous literature regarding music perception has identified how a small percentage of the population can perform at levels much lower than the general population at music perception tasks. This much reduced music perception ability has been named congenital amusia or 'tone deafness'. Kalmus and Fry (1980) stated that approximately 4% of the general population will exhibit congenital amusia. More recently there has been disagreement on the accuracy of this approximation (Henry and McAuley, 2010), nonetheless it appears that there are a certain number of people who are significantly worse at pitch perception and other music perception abilities compared to the general population.

It seems quite likely that the participant who was excluded from testing after the first session because they were unable to complete the tasks at the least difficult level, may have been exhibiting congenital amusia to some degree. It may also be the case that the two participants (8 and 9) who were removed at the beginning of data analysis may also have exhibited congenital amusia to a varying degree. However, participants 8 and 9 appeared to show music perception abilities at similar levels to the other participants for some conditions and not others. This implies that there may have been another reason why they were not performing well in some conditions. It is possible that the participants were not concentrating on the tasks at a consistent level throughout all the test sessions which produced this variability in the data.

The design of the present experiment involved participants completing six different music perception tests with no more than five minutes between each task for each session. This meant that the testing was quite lengthy and tedious for the participant. During the experiment some participants reported that they found it difficult to concentrate because they had to sit through repeated exposure to the same tests. Observation of the order of testing for individual participants shows that the tests with the most extreme outliers were presented to the participant towards the end of a test session. This further indicates that it may have been a concentration problem which contributed to the outliers rather than possible congenital amusia. The number of tests per session was a limitation of this study in terms of participant concentration and fatigue, however a repeated measures design was the best way to obtain a good measure of test-retest reliability. It is important to note that in a clinic situation, testing would not be repeated to this extent over this short time scale, therefore there are unlikely to be the same problems with fatigue and losses in attention.

It is clear that some participants in the normal hearing population may exhibit congenital amusia and perhaps at least one participant in the present study was unable to complete the tests provided in the SOECIC MTB possibly because of this. It is important to take this into account when the SOECIC MTB is being used for testing CI users, as there will be limited usefulness for someone with congenital amusia.

The SOECIC MTB has been primarily designed for CI users and therefore has a relatively long duration gap between each presentation of stimuli. Section 1.5 of this research project discusses how this may actually make the task more difficult for normal hearing listeners. The participants did in fact report that they found the task very slow moving in the present study and thought that differences in pitch would be more easily discriminated if the duration of silence between the three intervals was shorter. This was because the longer duration gap meant that it was more difficult to remember the three tones that had been presented. It was hypothesised that the 3I2AFC procedure with a reference tone may help counteract the long duration gaps and reduce the impact of the added cognitive component of using memory to remember all three stimuli presented. The results demonstrated that participants did in fact achieve smaller discrimination scores in the 3I2AFC procedure. However, once the 3I2AFC scores had been corrected to allow for the higher level of detectability required for the 3I3AFC procedure, participants did not perform differently for the two different procedures. A problem with this is that using the correction factor is making an assumption and may not account for the difference in results for the two procedures completely. Therefore it is possible that participants achieved lower discrimination scores using the 3I2AFC procedure not only because they had a higher chance of selecting the correct response but because the reference tone reduced participants' reliance on their ability to remember all three tones.

The selection of non-musicians for the present research project was based on the selection criteria laid out by Paynter (2010). This meant that the non-musician data collected for the present study could be compared to Paynter's (2010) musician and non-musician data. Although every effort was made to adhere to the selection criteria it was difficult to recruit participants who had received the same amount of musical training. Some participants reported having had no musical training, whereas others had received lessons in a musical instrument at school but did not partake in musical activity anymore. It is questionable where this arbitrary cut off point for musicians and non musicians should lie. It is clear that past research has shown musicians to perform at a higher level in music perception tasks compared to non-musicians (Micheyl et al., 2006). It is likely that different levels of music training could influence the effect it has on music perception. This will need to be taken into account when testing CI users and would be an interesting topic for further research. Also, the influence of CI users' duration of deafness prior to implantation on musicianship status is an important factor to investigate.

The present SOECIC MTB testing system implements a roving volume control for the pitch tests in order to eliminate any loudness cues participants may have used to identify the correct response. This varies the level of the stimuli in a range of 3dB (A) above and below the master volume setting. This appears to be effective at the easier levels of the test, however, as the pitch differences get smaller, the roving volume can become misleading and confuse the participant. This is because once the participant is close to threshold, the pitch differences are perceptually very small. Therefore, a 3dB(A) change in level could be mistaken as the interval that was different in pitch. Some participants reported that the change in the level of the stimuli was confusing when close to threshold. Paynter (2010) observed this previously and suggested that perhaps the amount that the volume roves should get adaptively smaller as the pitch differences get smaller.

## 4.5 Further research

There are a number of factors that have been highlighted in this chapter which would be interesting to investigate in future research. The SOECIC MTB is designed for CI users, therefore further research should assess if the same level of test-retest reliability can be demonstrated when testing CI users. Ideally, this study would have been carried out on a larger sample of participants so future research could investigate test-retest reliability in a larger group of normal hearing listeners. The presence of a response bias could be examined in terms of the present study’s data, to find out if there were any differences between the two different procedures. This could be used as another indicator towards the most effective response method for the SOECIC MTB. A number of other parameters could be investigated in terms of the test-retest reliability, for example, the number of reversals and the method of the adaptive procedure (such as, a two down one up versus a three down one up procedure). Another factor could be to research into the influence of different levels of musicianship on music perception ability.

## 4.6 Conclusion

There have been a number of limitations outlined in this chapter, for example, the repeated exposure to the same tests, the selection of non musicians, and the calculation of the identification scores. However despite these difficulties it is possible to conclude that this research project has shown that over three repeated test sessions, the SOECIC MTB exhibits test-retest reliability for a sample of 18 normal hearing listeners. Correlations showed that the 3I3AFC procedure appeared to produce more consistent results compared to the 3I2AFC procedure. This implies that the 3I3AFC format of the software should be maintained over the 3I2AFC procedure for future testing. A larger sample would have been desirable and is recommended to be investigated in the future in order to examine whether test-retest reliability still holds in a larger group and for CI users.

# 5 References

Brehm, J.W. and Self, E.A. (1989) The intensity of motivation. *Annual Review of Psychology,* 40, 109-31.

British Society of Audiology (2004) Recommended procedure: Pure tone air and bone conduction threshold audiometry with and without masking and determination of uncomforTable loudness levels.
Accessed 15th April, available from www.thebsa.org.uk/docs/RecPro/PTA.pdf.

Clark, G.M. (2006) The multiple-channel cochlear implant: the interface between sound and the central nervous system for hearing, speech, and language in deaf people-a personal perspective. *Philosophical Transactions of the Royal Society B,* 361, 791-810.

Cooper, W.B., Tobey, E. and Loizou, P.C. (2008) Music perception by cochlear implant and normal hearing listeners as measured by the Montreal battery for evaluation of amusia. *Ear and Hearing,* 29, 618-626.

Drennan, W.R. and Rubinstein, J.T. (2008) Music perception in cochlear implant users and its relationship with psychophysical capabilities. *Journal of Rehabilitation Research and Development,* 45(5), 779-790.

Field, A. (2005) Discovering statistics using SPSS (Second edition). SAGE Publications Limited, London.

Fitzgerald, D., Fitzgerald, H., Brockmeier, S.J., Searle, O., Grebenev, L., Nopp, P. (2006) Musical sounds in cochlear implants (Mu.S.I.C.) test. Innsbruck, MED-EL.

Gelfand, S. A. (1998) Hearing: An introduction to psychological and physiological acoustics (Third edition). Marcel Dekker, New York.

Gelfand, S. A. (2009). Essentials of audiology (Third Edition). Thieme, New York.

Gescheider, G.A. (1997) Psychophysics: the fundamentals (3rd Edition). Lawrence Erlbaum Associates, Mahwah, NJ.

Gfeller, K., Christ, A., Knutson, J.F., Witt, S., Murray, K.T. and Tyler, R.S. (2000) Musical backgrounds, listening habits, and aesthetic enjoyment of adult cochlear implant recipients. *Journal of American Academy of Audiology*, 11, 390-406.

Gfeller, K. and Lansing, C. (1992) Musical perception of cochlear implant users as measured by the primary measures of music audiation: an item analysis. *Journal of Music Therapy*, XXIX, 18-39.

Gfeller, K., Olszewski, C., Rychener, M., Sena, K., Knutson, J.F., Witt, S and Macpherson, B. (2005) Recognition of "real-world" musical excerpts by cochlear implant recipients and normal-hearing adults. *Ear and Hearing*, 26, 237-250.

Gfeller, K., Turner, C., Oleson, J., Zhang, X., Gantz, B., Froman, R., and Olzewski, C. (2007) Accuracy of cochlear implant recipients on pitch perception, melody recognition, and speech reception in noise. *Ear and Hearing*, 28, 412-423.

Gfeller, K., Turner, C., Mehr, M., Woodworth, G., Fearn, R., Knutson, J.F., Witt, S., and Stordahl, J. (2002) Recognition of familiar melodies by adults cochlear implant recipients and normal hearing adults. *Cochlear Implants International*, 3(1), 29-53.

Gfeller, K., Woodworth, G., Robin, D.A., Witt, S. and Knutson, J.F. (1997) Perception of rhythmic and sequential pitch patterns by normally hearing adults and adult cochlear implant users. *Ear and Hearing*, 18(3), 252-260.

Gordon, E.E. (1979) Primary Measures of Music Audiation. G.I.A. Publications, Chicago.

Kalmus, H., and Fry, D.B. (1980) On tune deafness (dysmelodia): Frequency, development, genetics and musical background. *Annals of Human Genetics*, 43, 369-382.

Kang, R., Nimmons, G.L., Drennan, W., Longnion, J., Ruffin, C., Nie, K., Won, J.H. Worman, T. Yueh, B., Rubinstein, J. (2009) Development and validation of the University of Washington clinical assessment of music perception test. *Ear and Hearing*, 30, 411-418.

Klein, S.A. (2001) Measuring, estimating and understanding the psychometric function: A commentary. *Perception and Psychophysics*, 63, 1421-1455.

Kong, Y.Y, Cruz, R., Ackland Jones, J. and Zeng, F.G. (2004) Music perception with temporal cues in acoustic and electric hearing. *Ear and Hearing*, 25, 173-185.

Laneau, J., Wouters, J., and Moonen, M. (2004) Relative contributions temporal and place pitch cues to fundamental frequency discrimination in cochlear implantees. *Journal of the Acoustical Society of America*, 116, 3606-3619.

Leal, M.C., Shin, Y.J., Laborde, M., Calmels, M., Verges, S., Lugardon, S., Andrieu, S., Deguine, O. and Fraysse, B. (2003) Music perception in adult cochlear implant recipients. *Acta Oto-Laryngologica*, 123(7), 826-835.

Leek, M.R. (2001) Adaptive procedures in psychophysical research. *Perception and Psychophysics,* 63 (8), 1279-1292.

Levit, H. (1979) Transformed up-down methods in psychoacoustics. *Journal of the Acoustical Society of America*, 49 (2), 467-477.

Looi, V. (2008) The effect of cochlear implantation on music perception. *Otorinolaryngol,* 58, 169-190.

Looi, V. and She, J. (2010) Music perception of cochlear implant users: A questionnaire, and its implications for a music training program. *International Journal of Audiology*, 49, 116-128.

McDermott, H.J. (2004) Music perception with cochlear implants: A review. *Trends in Amplification*, 8(2), 49-81.

Micheyl, C., Delhommeau, K., Perrot, X., and Oxenham, A.J. (2006) Influence of musical and psychoacoustical training on pitch discrimination. *Hearing Research*, 219, 3647.

National Institute for Health and Clinical Excellence (NICE) technology appraisal guidance 166 (2009) Cochlear implants for children and adults with severe to profound deafness. Accessed 12th April 2010, available from www.nice.org.uk/TA166.

Nimmons, G.L., Kang, R.S., Drennan,W.R., Longnion, J., Ruffin, C., Worman, T., Yueh, B. and Rubinstein, J.T. (2005) Clinical assessment of music perception in cochlear implant listeners. *Otology and Neurology*, 29, 149-155.

Paretz, I., Champod, A.S., and Hyde, K. (2003) Varieties of musical disorders: the Montreal battery of evaluation of amusia. *Annals of the New York Academy of Sciences,* 999, 58-75.

Paynter, K.R. (2010) Music tests for the deaf: pitch and rhythm perception of cochlear implant users, normal hearing listeners and the effects of musical training. University of Southampton MSc thesis.

Zeitlin, L.R. (1964) Frequency discrimination of pure and complex tones. *Journal of the Acoustical Society of America*, 36, 1027-1027

# 6 Appendices

## 6.1 Appendix 1: Participant consent form

**Exposure Number:** ...............

***Consent form to be completed by adult subjects taking part in an experiment***
*(Adults are 18 years of age or older.)*

**University of Southampton**
**Institute of Sound and Vibration Research**

Before completing this form, please read the list of contra-indications which has been provided by the experimenter on the reverse of this form.
This consent form applies to a subject volunteering to undergo an experiment for research purposes. The form is to be completed before the experiment commences.
I, ..........................................................................................................................
of .........................................................................................................................
(address or department)

consent to take part in: Pitch and rhythm discrimination testing consisting of 3 sessions lasting less than 1 hour each.
to be conducted by: Rachel Lamb
during the period: June to September 2010

---

The purpose and nature of this experiment have been explained to me. I understand that the investigation is to be carried out solely for the purposes of research. I am willing to act as a volunteer for that purpose on the understanding that I shall be entitled to withdraw this consent at any time, without giving any reasons for withdrawal. My replies to the above questions are correct to the best of my belief, and I understand that they will be treated by the experimenter as confidential.

Date: ................................... Signed: .........................................................................
(Volunteer subject)

I confirm that I have explained to the subject the purpose and nature of the investigation which has been approved by the Human Experimentation Safety and Ethics Committee.

Date: ................................... Signed: .........................................................................
(Researcher in charge of experiment)

**This form must be submitted to the Secretary of the Human Experimentation Safety and Ethics Committee on completion of the experiment.**

## 6.2 Appendix 2: Questions for participants

1. **Please give details if you have recently received treatment, or are currently undergoing treatment, for any of the conditions listed below:**

**Troublesome Tinnitus** .........................................................................................
…………………………………………………………………………………………

**Current Ear Disease** (e.g. persistent ear pain, ear infection or ear discharge)
.......................................................................................................................

**Other (please specify)** .........................................................................................
…………………………………………………………………………………………

2. **Have you been exposed to loud noises in the last 48 hours?**

3. **Are you musically trained in an instrument or voice? If so, to what level?**

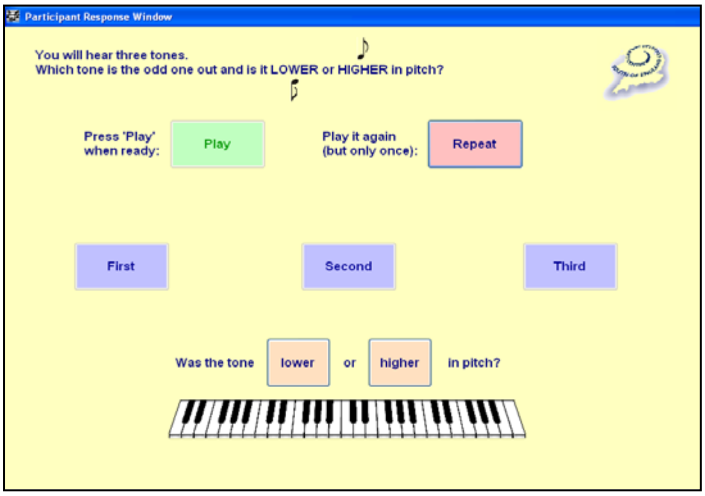
4. **Do you currently participate in musical activities or have you participated in musical activities in the last 5 years? Please give details.**

## 6.3 Appendix 3: Instructions for participants (3I3AFC procedure)

These tests will measure your ability to discriminate between small differences in pitch and rhythm.

### Pitch tests

You will hear three pure tones or piano tones. Two of the tones will be the same pitch whereas one will be different. You are required to click on the tone which is different. You will then be required to respond with whether the different note was higher or lower.

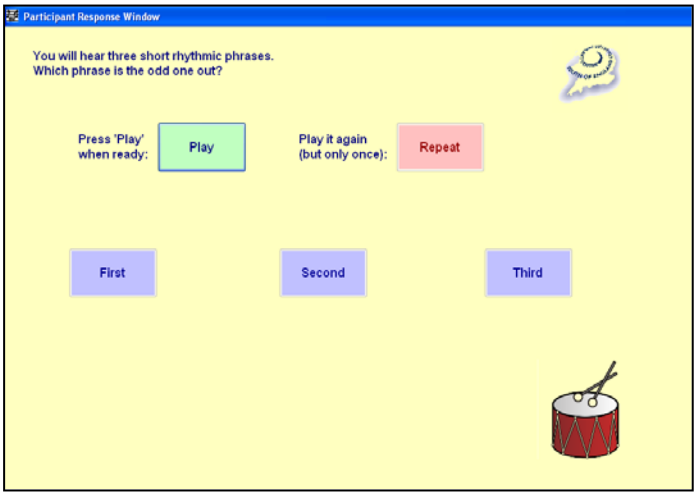


Participant Response Window

You will hear three tones.
Which tone is the odd one out and is it LOWER or HIGHER in pitch?

### Rhythm tests

You will hear three rhythms played one after the other. Two of the rhythms will be the same and one will be different. You are required to click on the rhythm which was different.



Participant Response Window

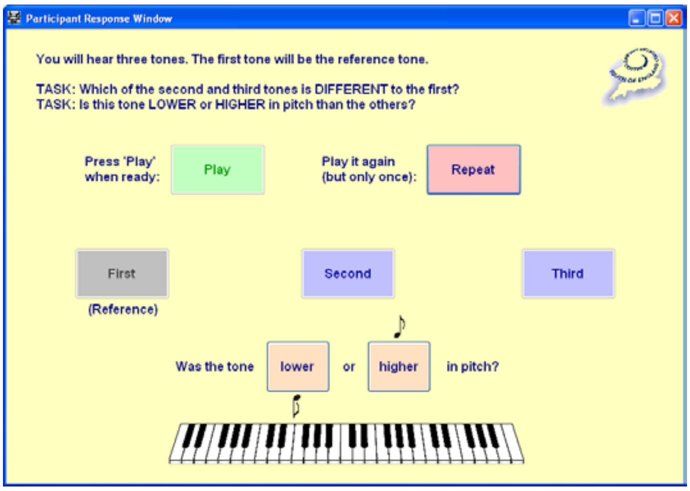
You will hear three short rhythmic phrases.
Which phrase is the odd one out?

## 6.4 Appendix 4: Instructions for participants (3I2AFC procedure)

These tests will measure your ability to discriminate between small differences in pitch and rhythm.

### Pitch tests

You will hear three pure tones or piano tones. The first tone will be a reference tone. You will then hear two more tones. One of these will be the same pitch as the reference tone and one will be different. You are required to click on the tone which is different. You will then be required to respond with whether the different note was higher or lower.

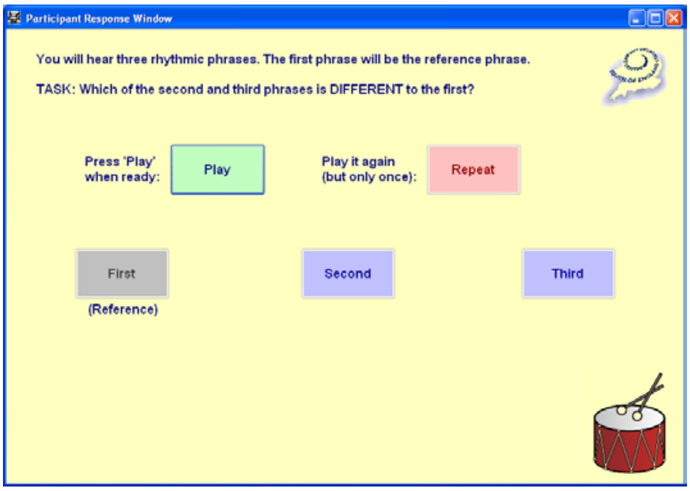


You will hear three tones. The first tone will be the reference tone.

TASK: Which of the second and third tones is DIFFERENT to the first?
TASK: Is this tone LOWER or HIGHER in pitch than the others?

### Rhythm tests

You will hear three rhythms played one after the other. The first rhythm will be a reference rhythm. You will then hear two more rhythms. One of the rhythms will be the same as the reference rhythm and one will be different. You are required to click on the rhythm which was different.



You will hear three rhythmic phrases. The first phrase will be the reference phrase.

TASK: Which of the second and third phrases is DIFFERENT to the first?

## 6.5 Appendix 5: CD-ROM containing experimental data

See attached CR-ROM: Data_Rachel_Lamb

- Folder structure: D:\Participant 1\Session 1
- Each Session folder contains txt files (participant responses for each trial) and bmp files (images of adaptive procedure) for the six conditions.
- The same folder structure applies for all participants (1-20) and sessions (1-3).