UNIVERSITY OF SOUTHAMPTON

FACULTY OF ENGINEERING AND THE ENVIRONMENT

Institute of Sound and Vibration Research

Volume 1 of 1

THE DEVELOPMENT AND EVALUATION OF A NEW TEST OF PITCH PERCEPTION FOR COCHLEAR IMPLANT USERS

by

Anne Marjorie Helen Wheatley

Thesis for the degree of Doctor of Philosophy

June 2018

This research is supported by an EPSRC CASE studentship with Cochlear Europe Ltd

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF ENGINEERING AND THE ENVIRONMENT

Institute of Sound and Vibration Research

Doctor of Philosophy

THE DEVELOPMENT AND EVALUATION OF A NEW TEST OF PITCH PERCEPTION FOR COCHLEAR IMPLANT USERS

by Anne Marjorie Helen Wheatley

Pitch perception, which is important for perceiving music, understanding tonal language and intonation cues and analysing the auditory scene, is limited by suboptimal temporal and spectral cues in cochlear implant (CI) users. The measurement of pitch perception has been quantified by several tests, however they demonstrate flaws in methodology and design and so the motivation behind this thesis was to improve on these tests.

Seven existing tests of pitch perception were evaluated using both normal hearing listeners (NHL) and CI users. The best performing test was the Melodic Contour Identification (MCI) test: it used a non-adaptive method, showed minimal floor and ceiling effects and had good reliability on retest. It had limitations too: the 9 melodic contours differed in their complexity, making some contours easier than others and the contours spanned up to 5 notes, meaning that it was not possible to assess single intervals.

A new test of pitch perception, the Pitch Contour Test (PCT) was designed to improve on existing tests. It used a non-adaptive method, which allowed the psychometric function relating pitch interval size to performance, to be estimated and visualised. The stimuli consisted of 4 contours, equal in difficulty and representative of single intervals, which allowed pitch discrimination and pitch ranking ability to be assessed simultaneously. It provided sufficient numbers of trials to ensure statistical confidence in the result and specified levels required for success.

The PCT was evaluated by comparing it to three existing tests of pitch perception: the University of Washington Clinical Assessment of Music Perception (UW CAMP), the Melodic Contour Identification (MCI) test and the South of England Cochlear Implant Centre Music Test Battery (SOECIC MTB), using NHL and CI users.

The PCT was superior to these tests with regard to numbers of trials, its ability to assess pitch discrimination and ranking simultaneously and its ability to estimate the psychometric function and be suitable for participants who demonstrate a non-monotonic function. The PCT performed similarly to the UW CAMP in terms of reliability, and similar to the MCI and the SOECIC MTB in terms of being sensitive to musicianship in the NHL group, however it did demonstrate floor and ceiling effects.

The development of the PCT impacts clinicians' ability to assess pitch perception in a more holistic way, and allows psychometric functions to be more fully explored. It has been used clinically to assess and determine the benefits to switching certain electrodes off in order to improve the listening experience for CI users. Use of the PCT throughout this thesis has demonstrated that both CI users and NHL can demonstrate non-monotonic psychometric functions. Using traditional adaptive methodologies when a psychometric function is non-monotonic can result in erroneous final results and without an estimation of the psychometric function, these results would appear to be accurate. Using a non-adaptive method when testing the pitch perception of CI users is therefore considered to be essential.

Table of Contents

Table of Tables

Table of Figures

Declaration of Authorship

I, Anne Marjorie Helen Wheatley declare that this thesis and the work presented in it are my own and has been generated by me as the result of my own original research.

THE DEVELOPMENT AND EVALUATION OF A NEW TEST OF PITCH PERCEPTION FOR COCHLEAR IMPLANT USERS

I confirm that:

1. This work was done wholly or mainly while in candidature for a research degree at this University;
2. Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
3. Where I have consulted the published work of others, this is always clearly attributed;
4. Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
5. I have acknowledged all main sources of help;
6. Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
7. Parts of this work have been published:

   Grasmeder, M L.; Verschuur, C; van Besouw, R M.; Wheatley, A; Newman, T A (2018) Measurement of Pitch Perception as a Function of Cochlear Implant Electrode and its Effect on Speech Perception with Different Frequency Allocations, *Int Journal of Audiology* 58(3)

Signed: .....

Date: .....

Acknowledgements

This piece of work is dedicated to my husband Seth and my beautiful boys Dylan and Jacob.

This has been a labour of love, inspiration, desperation and insanity at times over the last 8 years. It has not been easy completing this PhD, and I am indebted to so many people for allowing me time to work on it, especially over the last nine months.

Thank you To Seth for coping as I tried to switch my priorities from the boys to the PhD and back again several times a day. Thank you to my boys for adapting so well to being looked after by lots of other people and having to share Mama. Thank you to my Mum for being here week in, week out; thank you to Rose and Mart for your continued support. Thank you to my Dad and Lindsay for your support from afar.

Thank you so much to the kind hearted staff at the John McNeil Centre Opportunity Centre who provided me with the opportunity to start working again by looking after my precious boy, and continually supporting and cheering me on. Thank you to the staff at Julia's House Children's Hospice for looking after Dylan and allowing me a space to work. Thank you to Paisley from the Rainbow Trust for looking after my boys. And thank you to the team at Farley Nursery for having my tiniest tearaway and being so flexible and helpful.

Thank you to my endlessly supportive friends who have believed in me even when I have wanted to give up. Special thanks to my girls from uni who started their PhDs after me and finished before me, and have continued cheering me on from the side-lines both whilst we were in, and since coming home from hospital. Thank you to Sam for double Dadding with Seth for several weekends. Thank you to Elise and Louise for being there for me and sending me lovely things through the post. Thank you to Alexa for being a large part of why I am still sane.

I would also like to say thank you to my supervisors: Rachel, Mark, Daniel, Thomas and especially to Carl, for taking me on and being so understanding of my circumstances, and supporting me with voice recognition software over the years. Thank you to Jake from iSolutions for endlessly fixing huge computer problems. A huge thank you to Cochlear, for their financial support and travelling opportunities over the years. Thank you as well to Mary, for being an inspirational, kind and understanding colleague and mentor.

And finally thank you to my long-suffering cochlear implant and normal hearing volunteers who sat through an awful lot of musical notes to make this research possible.

List of Abbreviations and Definitions

AB	Advanced Bionics
ACE	Advanced Combination Encoder
AEP	auditory evoked potential
AFC	alternative forced choice
AM	amplitude modulation
BKB	Bamford-Kowel-Bench
cent	1/100 semitone
CF	centre frequency
CI	cochlear implant
ci	confidence interval
CIRGUS	Cochlear Implant Research Group University of Southampton
CIS	Continuous Interleaved Sampling
CNC	consonant nucleus consonant
CTC	Cochlear Technology Centre
CUNY	City University of New York
dB	decibel
df	degrees of freedom
DIS	discrimination
DL	difference limen
DR	dead region
E12	electrode 12

EAS	electroacoustic stimulation
ERAN	Early Right Anterior Negativity
ERP	event related potential
F0	fundamental frequency
F0Sync	F0 Synchronised ACE strategy
FFT	Fast Fourier Transform
FM	frequency modulation
FSP	Fine Structure Processing
HA	hearing aid
HINT	Hearing In Noise Test
HL	hearing loss
Hz	Hertz
ICC	intraclass correlation coefficient
ICC (A,1)	intraclass correlation coefficient: 2 way random model with single measures and absolute agreement
ICC (C,1)	intraclass correlation coefficient: 2 way random model with single measures and consistency
ICI	inter click intervals
IQR	Interquartile range
ISVR	Institute of Sound and Vibration Research
JND	just noticeable difference
kHz	Kilohertz
LPF	low pass filter
Max	maximum
MBEA	Montreal Battery for the Evaluation of Amusia

MCI	Melodic Contour Identification
MCIm	Melodic Contour Identification microtonal
MCS	method of constant stimuli
MDE	Modulation Depth Enhancement strategy
Mdn	Median
MEM	Multi-channel Envelope Modulation strategy
Microtonal notes	Pitch intervals less than 1 semitone
Min	minimum
MMN	mismatch negativity
Musician	defined as holding any musical qualification or self-reporting that they are a musician or both
NAP	neural activity pattern
NHL	normal hearing listener
NRES	National Research Ethics Service
PCT	Pitch Contour Test
PCTm	Pitch Contour Test microtonal
PDT	Pitch Discrimination Test
PDTi	Peak Derived Timing strategy
PMMA	Primary Measures of Music Audiation
RANK	Ranking
SAM	Sinusoidally Amplitude Modulated
SD	Standard Deviation
SG	spiral ganglion
SL	sensation level
SLM	sound level meter

SOECIC	South of England Cochlear Implant Centre
SPSS	Statistical Package for Social Sciences
SRT	Speech Reception Threshold
SSC	sung speech corpus
T and C levels	Threshold and comfort levels (that determine the comfortable dynamic range for electric stimulation of the cochlear implant)
T1/T2	Time 1/Time 2
TEPC	temporal envelope and periodicity cues
TFS	temporal fine structure
UK	United Kingdom
UOS	University of Southampton
USAIS	University of Southampton Auditory Implant Centre
UW CAMP	University of Washington Clinical Assessment of Music Perception

Chapter 1 Introduction

1.1 Background

Profound hearing loss (HL) is a worldwide problem and communication difficulties associated with it can be overcome by a cochlear implant (CI). In 2012 it was estimated that 324,200 people worldwide have received a CI (NIDCD Fact Sheet, 2016) and at the end of March 2017 there were 16,200 CI users within the United Kingdom (UK) (BCIG Annual Update, 2018). Whilst the perception of speech is excellent, the perception of music (Gfeller *et al.*, 2005; Olszewski *et al.*, 2005; Drennan and Rubinstein, 2008; Veekmans *et al.*, 2009) and appreciation of music (Mirza *et al.*, 2003; Philips *et al.*, 2012) is generally poor for CI users. This does not mean that CI users are not able to perceive and enjoy music; many do (Migirov, Kronenberg and Henkin, 2009; Philips *et al.*, 2012), and appreciation of music is not necessarily dependent on accuracy of perception (Gfeller *et al.*, 2000; Drennan *et al.*, 2015).

Reasons for the poor perception and appraisal of music are multifactorial, and can be broken down into problems with music, problems with cochlear implants, and problems with the profoundly deaf ear. The problem with music is that it contains a wide range of information in terms of frequency, dynamic range and simultaneous events, as well as having less redundancy than speech. The problem with cochlear implants is that they were originally designed to transmit speech, and speech and music have very different requirements for spectral resolution transmission. Electrode arrays are unable to stimulate every portion of the cochlea, and typically don't reach very far into the apex which should code for very low frequencies. Insertion of the electrode array is not easy and surgical trauma and electrode array kinking can complicate matters. Modern CIs are unable to transfer the temporal fine structure (TFS) of the incoming signal, due to the way that the envelope is extracted and the TFS is discarded, and they are not able to provide sufficient independent channels because they are limited by the number of electrodes in the array. The independence of these electrodes is dependent upon current requirements for each individual electrode and is affected by interactions between channels. Finally problems arise as a result of the profoundly deaf ear, with aetiology, auditory deprivation, and poor spiral ganglion (SG) survival all influencing success with a CI.

Chapter 1

Measures of music perception show significantly poorer scores compared to normal hearing listeners (NHL) (Gfeller *et al.*, 2005; Olszewski *et al.*, 2005; Drennan and Rubinstein, 2008; Veekmans *et al.*, 2009)(Gfeller *et al.*, 2005; Olszewski *et al.*, 2005; Drennan and Rubinstein, 2008; Veekmans *et al.*, 2009) and some CI users report disappointment, low satisfaction and reduced music listening after implantation compared to pre-deafness (Gfeller *et al.*, 2000; Mirza *et al.*, 2003; Lassaletta *et al.*, 2007; Looi and She, 2010). Generally CI users perform at a slightly poorer level than NHL on rhythm-based tasks, however melody, timbre and pitch perception-based tasks are substantially poorer than NHL (Gfeller and Lansing, 1992; Cullington and Zeng, 2010). In addition to its impact upon music, poor pitch perception can also impede tonal language comprehension (McDermott, 2004; Wang, Zhou and Xu, 2011) and perceptual segregation of simultaneous sounds (Oxenham, 2008).

Attempts to improve music perception for CI users include the development of new sound processing strategies, which have reports of mixed success, and music training programs, which often report success (Chen *et al.*, 2010; Cheng *et al.*, 2018). In order to successfully evaluate these methods, a valid assessment tool is necessary. Since the early 1990s, attempts to assess music perception in CI users have utilised existing measures (Gfeller and Lansing, 1991, 1992) and later, custom developments were published (Galvin, Fu and Nogaki, 2007; Nimmons *et al.*, 2008; Spitzer, Mancuso and Cheng, 2008; Brockmeier *et al.*, 2011b; van Besouw and Grasmeder, 2011). These assessment methods differ greatly in their approach, making test results difficult to compare from a research or from a clinical point of view. Methods of data collection and presentation vary, and details of test stimuli are often not easy to determine, thus complicating interpretation of results.

Limited validation assessment of these tests are reported in the literature (Kang *et al.*, 2009). Researchers are in need of appropriate tools to make comparisons between clinical choices or after intervention, and the tools available to them may not always be appropriate nor have sufficient information accompanying them in order to facilitate a suitable choice of measure.

A number of problems have been identified above: music perception for CI users is poor and motivation to improve this is high; it is essential that improvements to music perception are validly assessed; the existing assessment tools differ, are poorly validated, and it is not clear which tool(s) are most optimal and/or appropriate for the clinical environment. The aims of this research were to provide an informative and evaluative review of existing measures of pitch perception; to determine the suitability of these tests, and to attempt to improve on them with a new test of pitch perception for CI users.

1.2 Research questions

This thesis intends to answer the following general research questions (more detailed research questions are introduced at the start of Chapters 5 and 7):

1. Are existing pitch perception tests suitable and appropriate for use with CI users?

2. Of these existing pitch perception tests, does one (or more) test show greater performance than the others?

4. Is the Pitch Contour Test (PCT) an improvement on these existing pitch perception tests?

1.3 Original contributions

The original contributions contained within this thesis include:

1. An experimental and comparative review of 7 existing tests of pitch perception for CI users (Experiment I), which compared the methodological approaches of each test, their outputs, the methods by which results were calculated, and compared their final scores using CI users and NHL, assessed whether they were sensitive to musicianship and whether they were reliable on retest. This work highlights the vast differences between these tests which was not immediately obvious if these tests were to be used in the fast moving clinical environment. It highlights the importance of clarity and transparency regarding test design and calculation of results. This work is expected to be helpful for anyone who is using or plans to use music perception tests in a research or clinical environment with CI users, or anyone critically reviewing these tests within the literature.

2. The design of a new test of pitch perception for CI users: the PCT. The PCT is able to estimate the participants' ability to discriminate between pitches of different interval sizes, as well as simultaneously estimating their ability to pitch rank using different interval sizes. These results are recorded simultaneously, meaning that data is obtained for 2 separate tasks however the participant only does one task. Results are obtained for different frequencies and different timbres, allowing a more holistic approach to the assessment of pitch perception. The PCT uses the method of constant stimuli rather than an adaptive procedure and as such is able to estimate the shape of the psychometric function of pitch perception in CI users. The PCT contains sufficient repeated trials in order to provide

statistical confidence in the result; and provides clear instructions regarding the level needed to indicate success at a certain pitch interval.

3. An experiment (Experiment II) which evaluates the PCT using CI users and NHL. This work demonstrates the strength of the PCT in comparison to existing tests of pitch perception. It highlights the frequencies and timbres that demonstrate the most validity for use with CI users. It demonstrates the existence of non-monotonic psychometric curves both in CI users and in some non-musician NHL. This adds to the body of evidence indicating that some CI users demonstrate non-monotonic psychometric functions and as such that adaptive measures of pitch perception are not appropriate. This work also provides evidence that adaptive measures of pitch perception may also be inappropriate for some NHL.

1.4 Publications

The findings of Experiment I were presented:

- Orally at University of Gent, Belgium, March 2011
  Music perception tests for cochlear implant users

- Orally at University of Southampton: Music perception assessment and rehabilitation for cochlear implant users: an interactive afternoon, July 2011
  Music tests for cochlear implant users

- Orally at University of Southampton, Cochlear Implant Research Group University of Southampton (CIRGUS) talk, August 2011
  Higher or lower: does being a musician help determine pitch direction?

- Orally and in poster form at the second British Society of Audiology Conference and Short Papers Meeting 2011, in Nottingham, UK, 7 to 9 September 2011
  Wheatley, AMH., van Besouw, RM., & Lutman, M. 2011. Music perception tests for cochlear implant users: a review

- Orally at *'Hear the music of a cochlear implant & habilitation masterclass',* Cochlear Technology Centre, Mechelen, Belgium, November 2011

Music perception tests for cochlear implant users

- In poster format at The 12th International Conference on Cochlear Implant and Other Auditory Technologies (CI 2012), in Baltimore, Maryland USA, May 2012
  Wheatley, AMH., van Besouw, RM., & Lutman, M. 2012. An evaluation of pitch perception tests for cochlear implant users

- Orally at Friedberg Cochlear Implant Symposium, Friedberg, Germany, 22 June 2012
  An evaluation of pitch perception tests for cochlear implant users

- Orally and in poster form at the third British Society of Audiology Conference and Short Papers Meeting 2012, 5-7th September 2012
  Wheatley, AMH., van Besouw, RM., & Lutman, M. 2012. An evaluation of pitch perception tests for cochlear implant users

- Orally at *'Hear the music of a cochlear implant & habilitation masterclass',* Cochlear Technology Centre, Mechelen, Belgium, October 2012
  An evaluation of pitch perception tests for cochlear implant users

The findings from Experiment II were presented:

- In poster format at the Human Sciences Group Meeting, University of Southampton, July 2013
  Developing a test of pitch perception for cochlear implant users

- Orally at *'Hear the music of a cochlear implant & habilitation masterclass',* Cochlear Technology Centre, Mechelen, Belgium, October 2013
  Assessment of music perception for cochlear implant users

- In poster form at the British Academy of Audiology Annual Conference 2013, Manchester UK, Nov 2013
  Wheatley, AMH., van Besouw, RM., Rowan, D & Stainsby, T 2013. A Pitch Contour Test for cochlear implant users

- The PCT has also been used in Mary Grasmeder's thesis and this work has been published:
  Grasmeder, M L.; Verschuur, C; van Besouw, R M.; Wheatley, A; Newman, T A (2018)

Measurement of pitch perception as a function of cochlear implant electrode and its effect on speech perception with different frequency allocations, Int Journal of Audiology 58 (3)

1.5 Thesis structure

Chapter 2 provides an overview of pitch perception, including discussion of ways to quantify pitch, and presents the evidence surrounding the various pitch perception models.

Chapter 3 introduces the CI and describes the effect that processing has on the perception of pitch; the improvements for pitch perception and the pitch perception ability of CI users.

Chapter 4 details the methods available to measure pitch perception in CI users, presents the pitch perception tests investigated in this thesis and outlines the ideal test qualities required in a pitch test for CI users.

Chapter 5 introduces Experiment I, where existing tests of pitch perception are reviewed and evaluated using NHL and CI users. Each test is critically assessed with both NHL and CI users, according to the ideal test criteria presented in Chapter 4.

Chapter 6 explains the development of the new Pitch Contour Test, and justifies the decisions behind each of its features.

Chapter 7 introduces Experiment II, where the PCT is evaluated against the University of Washington Clinical Assessment of Music Perception (UW CAMP) test and the Melodic Contour Identification (MCI) test with CI users, and the South of England Cochlear Implant Centre Music Test Battery Pitch Discrimination Test (SOECIC MTB PDT) and the MCI with NHL.

Chapter 8 provides a brief overriding discussion addressing the main research questions in the thesis, discusses the limitations of the studies and makes suggestions for future improvements of the PCT.

Chapter 2 Pitch perception

The ability to successfully hear the pitch of sounds enables musical melodies to be followed and recognised, but even more importantly, pitch is essential for speaker identification (Zeng *et al.*, 2005), tonal language comprehension(McDermott, 2004; Wong *et al.*, 2008) and intonation. In addition it is essential for segregating sound sources and auditory scene analysis (Assmann and Summerfield, 1990).

2.1 What is pitch?

Pitch is the psychological correlate of frequency, and as it is a perceptual construct, it cannot be measured directly (Moore, 2003). 'Pitch [is] that attribute of auditory sensation in terms of which sounds may be ordered on a scale extending from low to high. Pitch depends primarily on the frequency content of the sound stimulus, but it also depends on the sound pressure and the waveform of the stimulus' (ANSI, 1994). A more musical definition is 'pitch is the perceptual attribute of a sound that can be used to produce melodies (Plack and Oxenham, 2005). Because pitch is a perceptual phenomenon, it cannot be related to the stimulus alone, but also depends upon the transformation of the stimulus that occurs by passing through the auditory system (Yost, 2009).

Intensity affects the relationship between frequency and pitch. Stevens, (1935), using 3 observers, found that generally, high frequencies increased in pitch with increasing intensity, and that low frequencies decreased in pitch with increased intensity. Intensity had little to no effect at mid frequencies e.g. 1000 Hertz (Hz). Verschuure & Van Meeteren (1975) showed that different pitches can sound as though they have the same pitch due to a difference in the intensity. At 300 Hz, they showed a maximal change of 15 Hz, which amounts to approximately 60 cents, as the intensity changed from 30-70 decibel sensation level (dB SL). At 500 Hz, they showed a maximal change of 13 Hz, which amounts to approximately 50 cents, as the intensity changed from 30-70 dB SL. At 1000 Hz, they showed a maximal change of 15 Hz, which amounts to approximately 30 cents, as the intensity changed from 60-90 dB SL. They concluded that for groups of individuals, Stevens' Law (above) is generally followed, but when looking at individual responses, pitch changes could go up, down or be non-monotonic with alterations in intensity.

2.2 Musical pitch

2.2.1 Notes and their frequency

The relationship between pitch and frequency means that pitch is often referred to using the frequency scale in Hz. Moore (2012, p. 203) states that 'assigning a pitch value to a sound is generally understood to mean specifying the frequency of a pure tone having the same subjective sound as the pitch'. Figure 2.1 below shows the notes from a traditional 88 key piano, and their note names, ranging from A0 (27.5 Hz) to C8 (4186.01 Hz).

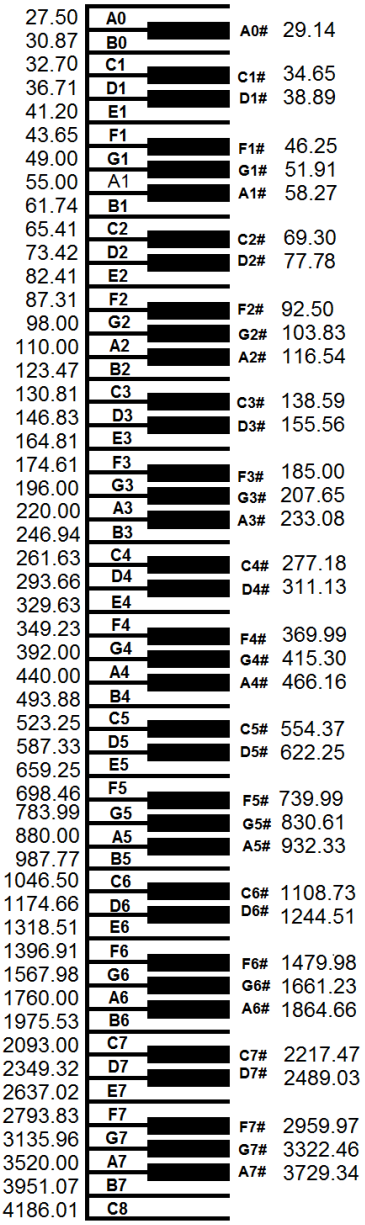


Figure 2.1 Piano note frequencies, calculated from A = 440 Hz, using the 12th root of 2, to 8 d.p. (1.05946309)

In Western music, the equal tempered 12 note chromatic scale is traditionally used, using concert pitch, and the note frequencies and their names on the standard 88 note piano are shown in Figure 2.1. The reference note for concert pitch is A4 = 440 Hz. Doubling the frequency results in A5 = 880 Hz, which is one octave higher and shares the same note name. The frequency of each note is calculated in reference to A4. The equal temperament means that each of the 12 semitones within the octave are equally spaced logarithmically and have equal ratios, and each semitone is equally divided into 100 cents. This means that there are 1200 cents in the octave. The ratio needed to calculate the next semitone step size can be found using the 12^{th} root of 2 = 1.05946309, so if A4 = 440 Hz, A#4 = 466.16 Hz. This can also be used to calculate the step size of 1 cent, using the 1200^{th} root of 2 = 1.00057779. Therefore, if A4 = 440 Hz, then A4+1(cent) = 440.26 Hz.

Pitch can be described as having two components; pitch height and pitch quality, or chroma, 'it's C-ness or D-ness' (Bachem, 1950). Evidence for these two being separate phenomena include that for frequencies above the typical region for 'musical pitch' e.g. above 4186 Hz (frequency of note C8, in Figure 2.1 above), chroma disappears, however pitch height can still be commented upon. Bachem also describes the difficulty in comparing the pitch height of notes from several instruments, with errors of 1-2 octaves occurring, however being able to determine simply if they are higher or lower, is much easier. In an earlier study, (Bachem, 1937, from (Bachem, 1950), used NHL with absolute pitch (the ability to successfully name a note correctly with no reference tone) to name notes and found that above 4000 Hz they started making errors: the chroma appeared too low. After 5000 Hz, all tones seemed to have the same chroma, and he visualised this in Figure 2.2 below. 'While the chroma reached a definite limit, tone height still increased, but in a peculiar manner; increased frequency resulted more in a weakening or peculiar thinning out of tones than in a greater tone height' (Bachem, 1950, pp 83).

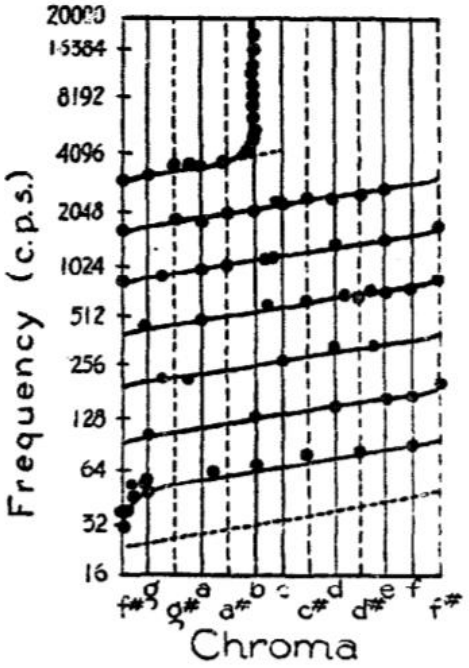


Figure 2.2 A graphical representation of chroma, from Bachem, 1950, reproduced with permission. This figure shows the cyclical nature of the notes (shown on the x axis) as the frequency increases, however at around 4100 Hz, this chroma, is lost, and as the frequencies increase, the note can be heard to get higher in pitch, but no interval or note can be heard.

As well as hearing each note within a melody, listeners also hear each interval and how they relate to each other – which is called the relative pitch. How the notes relate to each other can be described as the contour, and when listening to unfamiliar melodies, listeners are good at telling whether the contour is the same, regardless of whether it has been transposed. They are much poorer at determining whether the intervals have been preserved correctly. When presented with a task to compare original and distorted contour versions of familiar folk tunes, performance was very good (Dowling and Fujitani, 1971).

2.3 Understanding pitch perception

Early theories of pitch perception are the early theories of sound and hearing (Yost, 2009, de Cheveigne, 2005, Moore, 2012). Initial theories that followed the spectral approach were limited by the problem that the basilar membrane did not have the sensitivity to resolve all the possible pitches that the human ear can hear. Temporal theories were limited by the fact that the fastest

neuronal response could not code for the highest frequencies. A detailed account of the early theories will not be discussed here, however a number of notable phenomenon will be presented to indicate how the current theories developed.

2.3.1 The missing fundamental

Ohm's acoustic law, which stated that a Fourier type analysis was performed by the basilar membrane on the incoming sound, lead to the theory that the pitch was determined by the fundamental frequency, or the most prominent formant (from Yost, 2009). Removing the fundamental frequency from a harmonic tone caused problems for this theory, as even when no energy was present at the fundamental, a tone corresponding to that of the missing fundamental was still heard.

2.3.2 Unresolved harmonics and the residue

The cochlea can be modelled using 'auditory filters' which are described by Schouten, Ritsma and Lopes Cardozo, (1962) as a set of receptors, each of which respond to any particular frequency, and are limited by the width of the excitation curves. 'If the distance of two or more sinusoidal components is small compared to the width of the curves, these will materially overlap and hence the receptors in that region will respond to several frequencies at the time' (pp 998). At the apical end, which responds best to low frequencies, the filters are narrow and sharply tuned, and harmonic tones in this region are resolved. At the basal end, which responds best to high frequencies, the filters are wider and flatter and harmonics are often unresolved. Schouten's definition of the residue was that the periodicity pitch is heard because of the unresolved harmonics only, and that these harmonic interactions lead to the periodicity pitch.

2.3.3 Pitch shift of the residue

If upper harmonics are presented to a listener with a frequency shift for each harmonic, a change in pitch is heard. If the upper harmonics of the 200 Hz harmonic series are shifted by 40 Hz each, e.g. 440, 640, 840, 1040, the frequency spacing between each component is 200 Hz, however the heard pitch is 208 Hz (example given is from Yost, 2009).

2.3.4 The pitch of noise

Noise can elicit pitch. Evidence that energy at a particular place in the spectral domain is not required for pitch came from Burns and Viemeister (1981) who showed that sinusoidally amplitude

modulated (SAM) noise could elicit weak pitch cues, but enough to allow melody recognition. They used SAM noise to demonstrate that both melodies and musical intervals could be recognised successfully at a level much greater than chance. They tested 3 NHL on melody recognition using bandpass noise and found abilities of significantly higher than chance. They also used 3 trained NHL musicians and found that they could recognise intervals using SAM noise for musical intervals of 1, 2 and 3 semitones. Performance was better than chance, although performance at 3 semitones was better than at 2 and at 1. They concluded that both melodies and intervals could be heard and performance was better than chance when temporal cues alone were used.

2.3.5 The autocorrelation model

Current theories of pitch are based upon the autocorrelation model (Licklider, 1951). Licklider stated that the stimulus basis of sound consisted of 2 parts: the frequency and the periodicity, and that neither should be disregarded. In addition to the cochlea being regarded as a wave filter that analyses frequencies in the spectral domain, he proposed that autocorrelation analysis was carried out entirely in the time domain by the neural part of the system. The autocorrelation function 'is simply a running accumulation of the recent values of the product of [any given neuron as at a moment in time] and the same function delayed by [some lag time]. It provides a progressive description of the periodicity of the discharges of neuron *ij*' (Licklider, 1951, pp 129). At a neuronal level, this can be envisioned by a delay chain of neurons, which by definition and nature of the refractory period add delay. When these neurons synapse (after some delay) with cells that have input both from real time plus the lag time (cell C1,2,3), the resulting spatial synaptic summation leads to multiplication of the input, plus allows comparisons to be made. In addition, the temporal synaptic summation allows for integration to occur, co-ordinating the signal and allowing a running autocorrelation to take place.

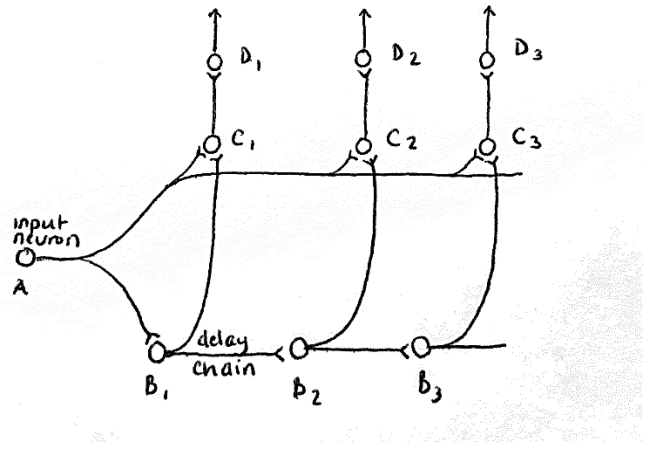


Figure 2.3 The basic schema of the neuronal autocorrelator (copied from Licklider, 1951). A: input neuron. B_1, B_2, B_3: delay chain neurons. C_1, C_2, C_3: delay chain neurons with excitation

created from the multiplication of A and B. D_1, D_2, D_3: represent the running integral and therefore the excitatory states display the running autocorrelation.

Figure 2.3 shows the initial input at A, going directly to C but also via the delayed neuronal chain B. The original plus the delayed signal are summed at C. This then feeds into D which fires at a rate proportional to the amount accumulated by the earlier points along the chain. Therefore, the output at D represents the running autocorrelation function of the signal.

Figure 2.4 shows that the input signal is now represented in two spatial dimensions: the frequency in the x direction, and the lag as caused by the delay chain of neurons (B1, B2, B3, Figure 2.3) in the lag direction, which is all happening in real time.

The summary autocorrelogram (Figure 2.4 B) shows that the output produces a high correlation activity at 5 ms, indicating that the input frequency was 200 Hz, and is shown to remain even if energy at 200 Hz does not occur.

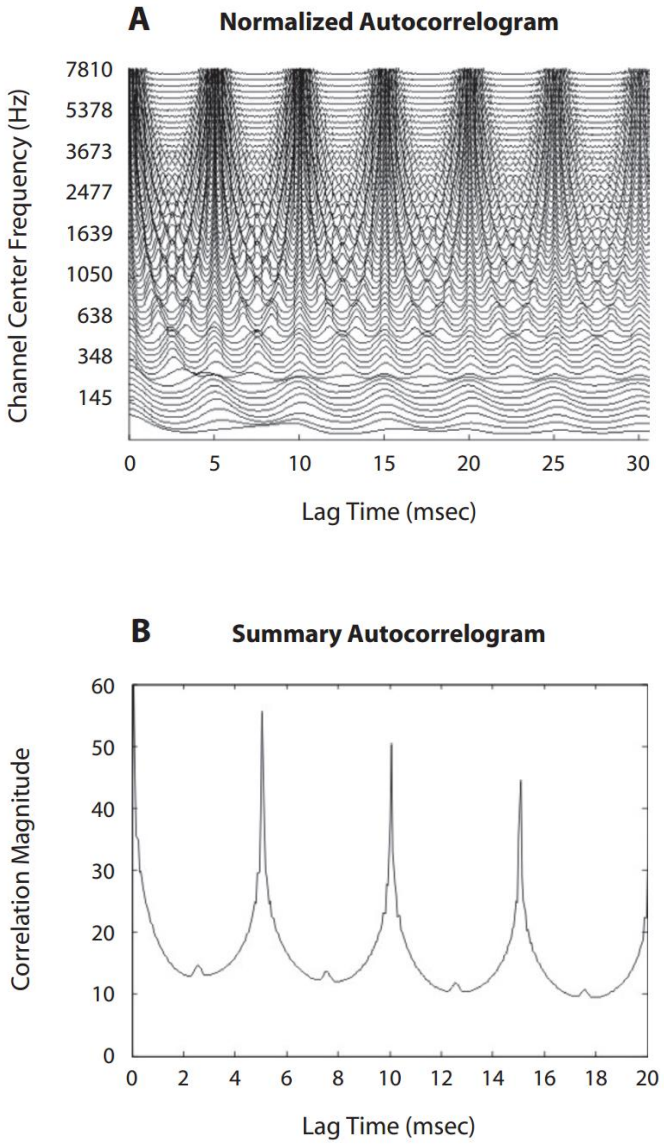


Figure 2.4 Autocorrelograms, from Yost, 2009 Reproduced with permission

2.3.6 Against autocorrelation

Autocorrelation cannot explain temporal asymmetry that is seen in the auditory system. Irino and Patterson, (1996) and Patterson and Irino (1998) have shown that using ramped and damped stimuli can elicit noise with a pitch saliency, however when a damped sinusoid or noise stimuli is reversed to create a ramped sine or noise stimuli, as seen in Figure 2.5 below, the saliency of the pitch is greater. Two forms of autocorrelation model were presented with ramped or damped stimuli and neither could explain the temporal asymmetry seen. The auditory image model adds to Licklider's 1951 autocorrelation model by including the strobed temporal integration component. This represents the auditory image in response the neural activity pattern (NAP). Figure 2.5 below depicts the output from the strobed temporal integration module described in Patterson and Irino (1998). This module takes the envelopes of the NAP and feeds them to the delta-gamma process,

which decides whether a 'strobe' (the output of the model) should be issued by applying a process of an adaptive threshold, an accumulator and the delta gamma process. The delta gamma is the derivative of the envelope of the NAP, and this process decides to issue a strobe when the accumulator output exceeds the adaptive threshold: and temporal integration is initiated.

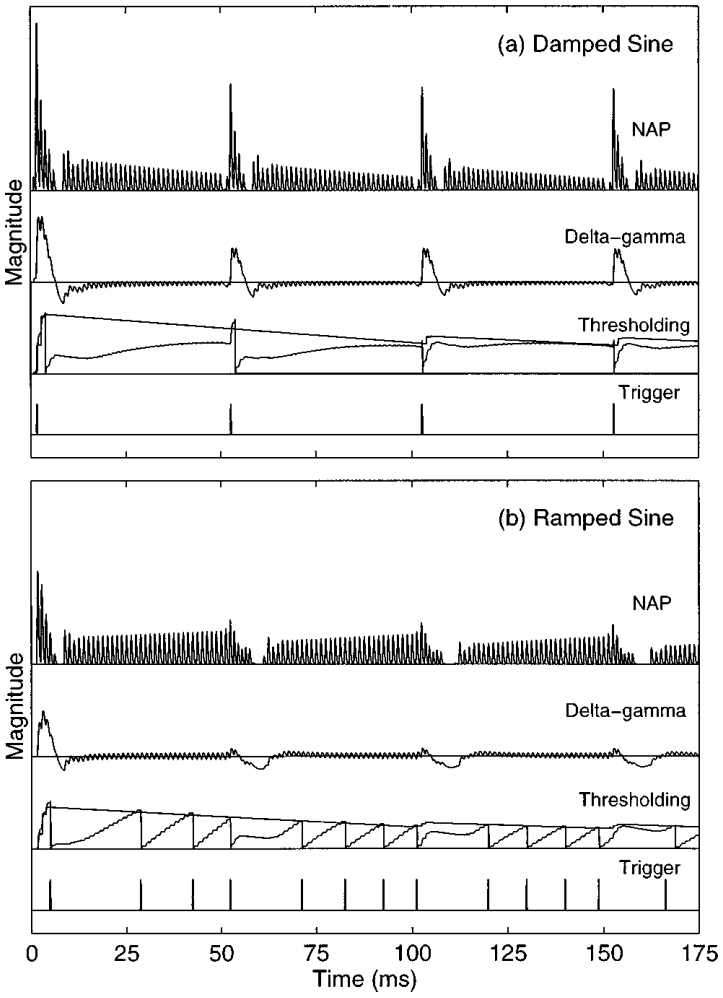


Figure 2.5 Damped and ramped sine tones, create a different pitch, from Patterson and Irino (1998), reproduced with permission.

Another way to elicit a weak pitch is using click trains to create a 'rattle pitch' (Kaernbach and Demany, 1998). Click trains that are high pass filtered above 6000 Hz and added to low pass noise to remove any possibility of spectral distortion products can be said to have a pitch percept due to the click rate. If the inter click intervals (ICI) are then interspersed with a random click (e.g. "kxx", "kxxx", where k = 10 ms and x = random interval), which interrupts the click train, the pitch saliency is lost. Attenuating the random clicks causes the pitch to return at an average of -9.2 dB. If the form "abx" is used (where a+b combined result in an interval of 10 ms, and "x" remains a random

interval), the AC model would predict a similar outcome – with a peak at 10 ms. However, perceptually, whist kxx tasks are performed well (around 80% correct), abx tasks are much harder to differentiate from random click trains and performance is much closer to chance. These results imply that no temporal information is conveyed by non-consecutive neural spikes.

2.4 The effect of HL on pitch perception

Hearing loss has a great impact on the ability to hear pitch. The broadening of auditory filters means that frequency selectivity, the ability to hear out the components of an acoustic complex, is reduced. Smith *et al.*, (1987) showed that when treated with a drug that selectively destroyed the function of the OHC, but leaving the IHC intact, monkeys showed a reduction in hearing threshold and their psychophysical tuning curves became broader and flatter. Adults with sensorineural HL have been shown to perform much more poorly on tasks that require frequency selectivity (Bernstein and Oxenham, 2006). The loss of IHC and type 1 auditory nerve fibres doesn't always lead to a detriment in hearing in quiet, as thresholds can still be good under favourable conditions, however performance is much poorer in noise (Salvi *et al.*, 2017).

Chapter 3 Cochlear implantation

3.1 The cochlear implant

The CI is an implantable electronic device that delivers pulses of current to the auditory nerve in order to cause a sensation of hearing which can be interpreted as sound. Current UK guidelines define CI candidates as people with bilateral severe or profound deafness (average audiometric threshold at 2 and 4 kilohertz (kHz) of 90 dB HL or poorer) and a best aided Bamford-Kowal-Bench (BKB) sentence test score of ≤ 50% (National Institute for Health and Care Excellence, 2009). The typical location and positioning of the external and internal parts of the CI can be seen in Figure 3.1.

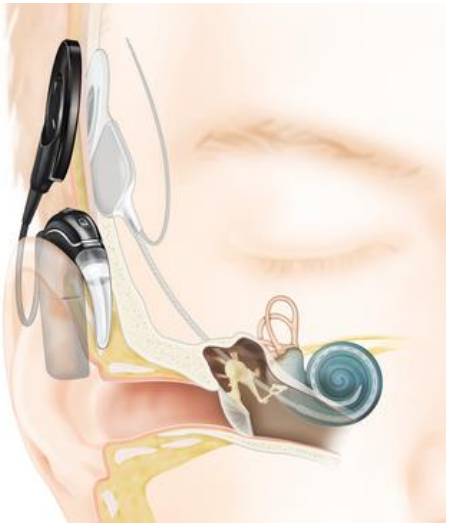


Figure 3.1. Diagram detailing where the CI external and internal devices are positioned on the head. *Image courtesy of Cochlear*

Cochlear implants consist of a microphone, a speech processor, a transmitter, a receiver-stimulator and an electrode array (see Figure 3.2). The acoustic input is transduced into a analogue signal by the microphone, and is then sent to the speech processor where the signal may be subject to pre-emphasis and boosting. The signal is then subjected to a filter bank of frequency analysers that divide the signal into several narrowband signals. The temporal envelope of this signal is also extracted and the TFS is discarded. This signal is converted to an electrical signal to send to the electrode array, which is arranged tonotopically: with high frequencies stimulated at basal parts of the cochlea and lower frequencies at more apical regions.

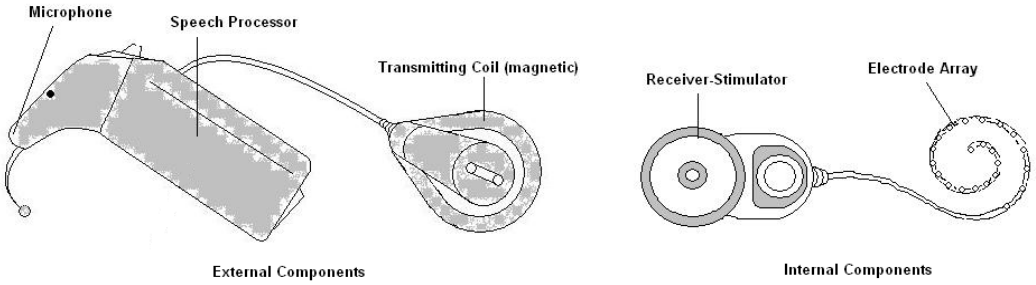


Figure 3.2. The components of a typical CI

3.2 The problem with pitch through a CI

3.2.1 Processing

In order to deliver sound to an array of up to 22 electrodes, the sound must be processed to extract the information that can be transmitted. Initially, the sound is transduced by the microphone from a pressure signal into an electrical signal. Once the sound has been picked up by the microphone, it undergoes a number of stages in information processing. Audio pre-processing then takes place consisting of amplification to a suitable level, pre-emphasis to boost high frequencies, automatic gain control, and automatic sensitivity control, to provide compression and a boost to quiet sounds, respectively. An anti-alias filter is also applied, cutting frequencies above 8000 Hz, so that a sampling frequency of 16000 Hz can be used. At this stage, the signal may be digitised by the analogue to digital converter, or the signal may be processed by a bank of band pass filters initially, and then digitised.

The next stage in information processing is frequency analysis and envelope extraction. There are a number of methods by which this can be achieved, including the use of band pass filters followed by half- or full-wave rectification and a low pass 'smoothing' filter, using the Fast Fourier Transform (FFT) with a weighted sum of the FFT bins to achieve the right number of channels, or using the Hilbert transform. The output from each of the frequency analysis filters is the slowly varying envelope. Rosen (1992) divided temporal information in speech into three parts: the temporal envelope, which consists of changes in time of less than 50 Hz; the periodicity, consisting of temporal changes ranging from 50 to 500 Hz; and the fine structure, consisting of temporal changes at a rate higher than 500 Hz, and typically ranging from 600 – 10,000Hz. Later studies (e.g. Ping *et al.*, 2010) have grouped the envelope and periodicity cues together for simplicity in CI research, referring to them as temporal envelope and periodicity cues (TEPC). This may make

more sense as temporal fluctuations of greater than 50 Hz are transmitted by the CI, meaning that periodicity cues as defined by Rosen (1992) are transmitted.

Envelopes are then used to amplitude modulate (AM) biphasic pulses for each channel (Loizou, 2006). These pulses are then delivered to the electrode within the boundaries of the maximum and minimum current which is set clinically when the patients' threshold and comfort levels are determined. These biphasic pulses are therefore carrying the envelope information in the form of AM. Current speech processing strategies either select 10-12 maxima e.g. the Advance Combination Encoder (ACE) strategy, or use fixed filters e.g. the Continuous Interleaved Sampling (CIS) strategy.

3.2.2 Envelope and temporal fine structure cues

The output from each frequency band is split into the temporal envelope and the TFS, the envelope is used to modulate the fixed pulse rate of the CI, and the TFS is discarded (Loizou, 2006), as CI users are generally unable to perceive temporal information above 300 Hz (McKay *et al*., 1994 and Wilson *et al.,* 2004). Envelope extraction (and subsequent TFS removal) is either achieved through half or full wave rectification and low pass filtering (LPF), or using the Hilbert transform. Both the ACE and CIS processing strategies use envelope cut off frequencies of 200 to 400 Hz (Arora *et al*., 2012, Loizou, 2006). Both methods result in the CI transmitting temporal information of greater than 50Hz, and therefore CIs are able to transmit more temporal information than just the temporal envelope as defined by Rosen (1992), and contain temporal periodicity cues as well. Therefore, within CI processing, the 'envelope cues' are typically referring to both envelop and periodicity cues, and can be better described as being temporal envelope and periodicity cues (TEPC). Rosen's (1992) definition of the TFS is from 500 or 600 Hz to 1000 Hz, and so current clinical CIs, with envelope cut off frequencies of 200-400 Hz are not able to transmit TFS. Temporal fine structure is important for timbre, voice quality and place (Rosen, 1992) and has been shown to be essential for pitch and lexical tone perception (Xu and Pfingst, 2003).

The varying importance of the TEPC and TFS cues for speech and pitch perception have been demonstrated using 'auditory chimeras' (Smith *et al.*, 2002). Auditory chimeras are created by processing two signals, and dividing them into 1-64 band pass filters, from 80 – 8820 Hz. These signals are then divided into their 'envelope' (which is actually the TEPC, as the Hilbert transform was used) and their TFS. The envelope from one signal is combined with the TFS of another in order to investigate the relative importance of each, by determining what is perceived by the listener.

They found that envelope cues dominated speech-speech chimeras when the number of channels was 4 or more, and therefore showed that the envelope information in speech is resistant to conflicting TFS from another sentence. When melody-melody chimeras were used, it was the TFS that dominated what melody was heard, up until 32 bands. At 48 and 64 bands, the melody with the envelope cues was heard more often, however Smith *et al*., (2002) report that both melodies were often heard. The fact that this crossover between the importance of envelope and TFS cues is so much higher for melody recognition (approximately 40 bands compared to 2), indicates the essential nature of TFS for melody recognition. Their findings indicate that the delivery of TFS via the CI, and in a way that CI users can access it (Wilson *et al*., 2004) is likely to improve melody and pitch perception, tonal language perception, access to interaural time difference cues for localisation, and improve speech understanding in noise.

The lack of TFS being transmitted through the CI is likely to be responsible for the problems surrounding pitch perception in CI users. In clinically available CI devices, up to 400 Hz of temporal information is typically available, and arguably, only up to 300 Hz of this may be accessible to the CI user (McKay *et al*., 1994). Kong *et al*., (2008) have shown that some CI users may be able to access pitch information from temporal rates of greater than 300 Hz, and some up to 500 Hz, and that this is dependent on the individual. This has led to the suggestion that time and effort might be better spent in determining suitable candidates for processing strategies that deliver more TFS, rather than trying to develop broader strategies to deliver greater TFS to all CI users.

3.2.3 **What is the effect of moving across channels in frequency?**

If there were no temporal cues available to CI users, and they had to rely solely on place cues to hear pitch, then a smooth frequency sweep would sound like a selection of tones based upon the centre frequency (CF) closest to the note at that moment in time, jumping from one CF to the next. Indeed, this is reported by some CI users. Describing a pure tone sweep which can be controlled by a dial, Barry Jacobson, who is a Med-El CI device wearer reports 'If the dial is turned still further, the tone will shift into another CI channel, but this will manifest as a discrete jump, rather than a continuous variation' (Jacobson, 2014). This would not only affect musical interval relationships which do not coincide with the CFs used by CIs, but would also affect the relationships between the harmonics of the notes, rendering that no musical note or any musical interval could ever sound as it was intended.

Another CI user reports hearing the note C6 (1046.5 Hz) as G6 (1568 Hz), and reports that notes higher than middle C (C4, 261.63 Hz) are not as distinct from each other as the notes below middle

C (Reed, 2016). This CI user also reports that when listening to a smooth frequency sweep, he perceives the lowest part of the sweep as one note, and then the notes rise smoothly for the two octaves below middle C. As the notes move above middle C, he reports that some notes sound as though they are missing, and the change between notes becomes less smooth (R. Reed, personal communication, 13th March 2019).

If it was the case that (all) CI users could only hear in discrete CF jumps, then it wouldn't be possible for them to hear intervals any smaller than the distance between two CFs. Below are some example CFs from the Cochlear CI, and their corresponding closest musical notes:

Electrode 22	250 Hz	B3	246.94 Hz
Electrode 21	375 Hz	F#4	369.99 Hz
Electrode 20	500 Hz	B4	493.88 Hz
Electrode 19	625 Hz	D#5	622.25 Hz
Electrode 18	750 Hz	F#5	739.99 Hz
Electrode 17	875 Hz	G#5	830.61 Hz
Electrode 16	1000 Hz	B5	987.77 Hz

The distance between electrode 21 and 22 is 7 semitones, between 20 and 21 is 5 semitones, between 19 and 20 is 4 semitones, between 18 and 19 is 3 semitones, between 17 and 18 is 2 semitones and between 16 and 17 is 3 semitones. Whilst pitch discrimination difference limens often fall around these indicated interval sizes, and average discrimination ability is in this region for CI users, there are CI users that are able to perform much better than this at both pitch discrimination and pitch ranking tasks (e.g. Gfeller *et al.*, 2002; Nimmons *et al.*, 2008; Jung *et al.*, 2009; Kang *et al.*, 2009; van Besouw and Grasmeder, 2011; Maarefvand, Marozeau and Blamey, 2013; Drennan *et al.*, 2015). This would indicate that the limited temporal pitch cues that are available, in the form of temporal fluctuations of up to 200-400 Hz depending on the envelope cut off frequency, are able to be used in addition to the crude place cues provided by the electrode CF. It has been proposed by Kong et al (2008) that rather than attempt to modify the stimulus to improve temporal processing, signal processing strategies should be matched to the individual CI user's temporal processing capabilities, in order to maximise the accessibility and functionality of temporal cues that are available.

3.2.4 Hardware

The hardware of the CI is also a limiting factor. Place pitch, defined as the perception of pitch which is heard as a result of the exact place of stimulation, is limited by the length of the inserted electrode array. Modern multichannel cochlear implants typically have 12 to 22 electrodes (Landsberger and Srinivasan, 2009), which are able to stimulate in response to frequencies from 200 to 8000 Hz (Svirsky, 2017). Stimulation of electrodes from apex (low frequency) to base (high frequency) of the cochlea typically demonstrate a tonotopic arrangement (Townshend *et al.*, 1987; McDermott and McKay, 1997). The ability of the cochlear implant to deliver successful place pitch cues depends on the independence of each electrode, without which the benefits of a multichannel implant are redundant (Townshend *et al.*, 1987).

Only being able to use up to 22 channels in place of 16,000 hair cells means that much spectral representation is lost. Loizou, Dorman and Tu, (1999) conducted a study using NHL listening to vocoder simulation and discovered the best performance in sentence testing was with 5 channels, achieving average scores of around 90%, whereas 2 channels was around 25%. As channels increased above 5, performance stayed around the same. Number of usable channels is determined by channel interaction. Crew, Galvin and Fu, (2012) used 20 NHL using CI simulations, and asked them to complete the MCI. The simulations included channel interaction by adding neighbouring temporal envelope information to nearby channels, to differing degrees to simulate no channel interaction, slight, moderate or severe. Performance was reduced as channel interaction was increased.

The common ground configuration occurs when current flows from the stimulating electrode to all other electrodes which are connected together to act as the ground. Monopolar stimulation occurs when the stimulating electrode discharges current to a distant electrode outside the cochlea, bipolar stimulation occurs when the current flows between two electrodes and tripolar stimulation occurs when the current flows between three cochlear electrodes and a ground electrode (Figure 3.3).

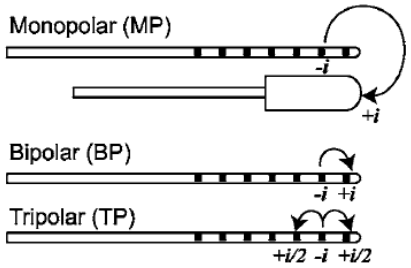


Figure 3.3 Monopolar, bipolar and tripolar electrode configuration, from Bierer (2007), reproduced with permission.

Busby *et al.*, (1994) conducted a comparison study with nine CI users and compared common ground, monopolar (returning pathway was the most basal electrode) and bipolar configurations (BP +1 for 8/9 and BP +2 for 1/9). Generally, tonotopic patterns were seen in all patients and all configurations, although for 3/9 patients, pitch reversals were reported for mid frequency electrodes with the common ground configuration. Whilst they report that significant differences in pitch estimates are likely to occur with the use of different electrode configurations, for approximately half of the patients they tested, no clear advantage existed between bipolar and common ground. Whilst monopolar stimulation is likely to cause wider current spread, a more uniform dynamic range and regular threshold and comfort levels were reported for the monopolar stimulation, making them easier to measure and indicating that this may be useful in difficult to test patients. Fielden *et al.*, (2015) used 8 CI users and a sung vowel test testing 6 and 3 semitones, and found no advantage of using tripolar stimulation over monopolar stimulation on this task.

Insertion depth and angle of insertions have both been shown to affect performance on speech tests. Yukawa *et al.*, (2004) showed that these factors significantly correlated with Consonant Nucleus Consonant (CNC) words, CNC phonemes and City University of New York (CUNY) sentence scores using 48 CI users. Greater insertion depth can also result in a greater range of independent frequencies perceived by the listener, as has been shown with 14 MedEl users, with a 31.5mm array. Insertion further into the apex (as achieved with deeper insertion) has indicated less perceptual differences between these electrodes (Landsberger *et al.*, 2014).

In addition, the electrically stimulated cochlea is not stimulated in the same region as an acoustically stimulated cochlea, there is mismatch (Figure 3.4). Speech recognition is optimised when electrical stimulation occurs close to the acoustic tonotopic cochlear location (Başkent and Shannon, 2004). A stimulus will often sound higher in pitch than it actually is (Loizou, 1998), resulting in perceptual change of speech and musical sounds, with familiar melodies sounding distorted (Swanson, 2008). Electric pitch perception has also been shown to change over time over the first few years of implantation (Reiss *et al.*, 2007). Spectral mismatch causes detrimental effects in speech perception, although these can at least in part be alleviated with training (Rosen, Faulkner and Wilkinson, 1999).

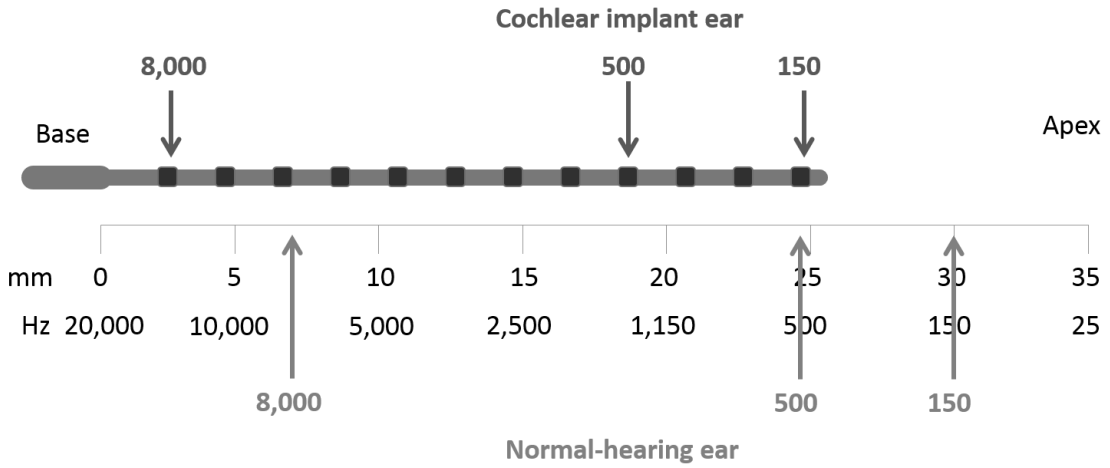


Figure 3.4 Diagram showing the potential spectral mismatch between the implant and the frequencies expected by the normal hearing ear, from Bernstein *et al.,* (2018), reproduced with permission

A recent summary of electrode array designs specifies that ideally, electrodes should be inserted atraumatically (e.g. aiming to preserve hearing if at all possible), should be placed as close to the modiolus wall (e.g. the central bony part of the cochlea), and the array should be fully inserted into the cochlea, as well as being easy for the surgeon to insert with as little complications as possible and be possible to explant if necessary (Dhanasingh and Jolly, 2017).

3.2.5 Human issues

Dead regions (DR) for people with cochlear HL are defined as loss of IHC and auditory nerve fibres, whereas in CI users the definition is particularly relating to are areas of local neural death, and these areas are likely to relate to longstanding hair cell loss. If DRs are stimulated with a CI (see

Figure 3.5), spectral information can become distorted (Won *et al.*, 2015), and current spread and channel interactions are likely (Macherey and Carlyon, 2014). If an electrode stimulates a DR within the cochlea, how that pitch is heard relates to how poorly functioning the surrounding areas are. Huss and Moore (2005) reported a listener who heard a 500 Hz tone that was presented to the DR as being around 3-4kHz, the area where their hearing was preserved, however another example was given where the tone heard was much closer to the original tone, and the temporal properties were more useful, thought to be due to the relatively good hearing thresholds immediately surrounding the DR.

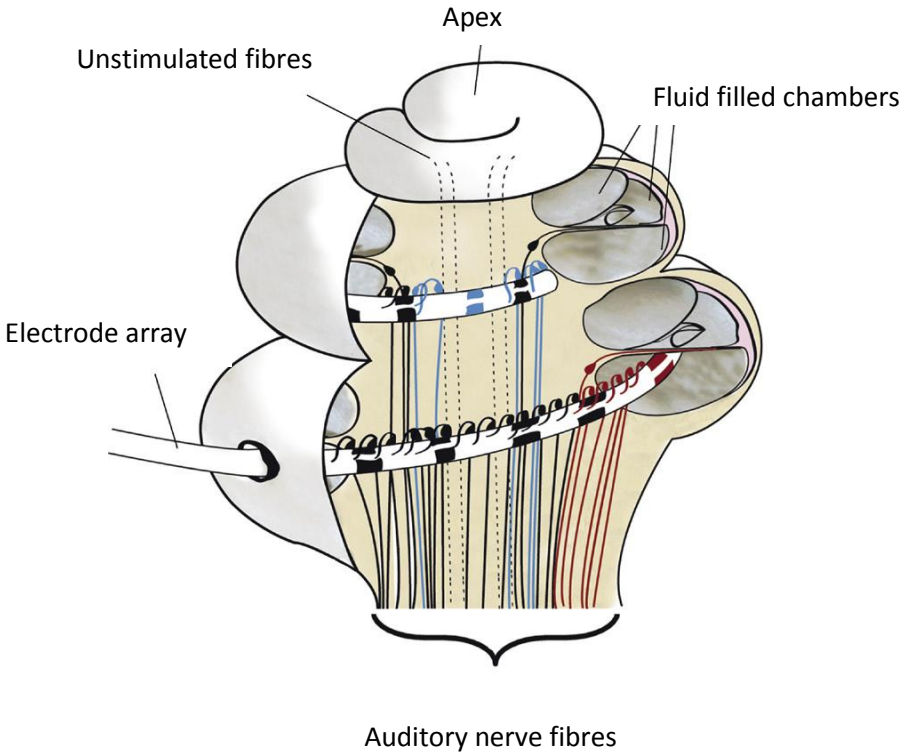


Figure 3.5 Dead regions within the cochlea. The central blue electrode with no fibres surrounding it lies within a DR, and current then spreads to neighbouring neurons that are already being stimulated by different electrodes. From Macherey and Carlyon (2014), reproduced with permission.

3.3 Attempts to improve the perception of pitch

3.3.1 Pitch processing strategy improvements

Experimental strategies have aimed to encode TFS in order to improve the perception of pitch in CI users. The Frequency Amplitude Modulation Encoding (FAME) strategy was proposed by Nie et al (2005) and aimed to encode both frequency modulation (FM) and amplitude modulation (AM), in order to be more like the normally functioning auditory nerve which does not use a fixed rate spike train but is able to fire synchronously with the stimulus waveform. Stickney *et al.* (2004) used the FAME with NHL using simulations and made comparisons between AM, or AM and FM simulations. They used a familiar melody test with rhythm cues removed, as well as speaker and vowel recognition task. They found improved scores for all NHL simulations with the FAME strategy, indicating the potential for transmitting TFS in this way.

Vandali *et al.* (2005) compared the Advanced Combination Encoder (which choses a number of maxima) and Combined Interleaved Sampling (which uses a fixed peak approach) with 4 experimental strategies: the Peak Derived Timing (PDTi), the Modulation Depth Enhancement (MDE), F0 Synchronized ACE (F0Sync) and Multi-channel Envelope Modulation (MEM). They used 11 CI users and tested them with a two alternative forced choice (2AFC) pitch ranking test, with intervals of 6 semitones, using male and female sung vowels. Whilst no specific benefits were seen when the data was analysed as a group, individuals gained benefit from different strategies. No pitch reversals were seen with the MDE, whereas they were seen for the F0Sync and ACE. The MDE, F0Sync and MEM were all significantly better than ACE for the male sung vowels at 6 semitones. Whilst all the new strategies coded F0 using deeper modulations compared to ACE, the MEM in addition coded the temporal peaks coincidentally across electrodes, minimising phase differences, and this is thought to be the reason for the benefits seen. A further study using MEM (Wong *et al.*, 2008) showed no improvements on a Cantonese hearing in noise test when compared to ACE, but there was an improvement compared to CIS, however ACE was rated higher than MEM by these 9 Cantonese speaking CI users.

Arnoldner *et al.* (2007) compared the Fine Structure Processing (FSP) with the CIS, using 14 CI users, and tested them using syllable, digit and sentence testing, as well as the rhythm, melody and number of instruments test from the MedEl MuSIC Test battery. They compared CIS scores at baseline with FSP at 12 weeks and found improvements in almost all speech tests, and in the rhythm and instrument test, but not the melody test. Magnusson (2011) however showed no significant improvement after upgrade from CIS+ to FSP. He used 20 newly implanted CI users who had been temporarily fitted with the CIS+ strategy, who were then upgraded to FSP, and tested them at 1 and 4 weeks, and 2 years later with speech tests in quiet and noise, and on quality of music.

The MP3000 strategy minimises the number of channels stimulated compared to ACE. The ACE strategy selects clusters of adjacent channels and this can result in interactions and forward masking, and so the MP3000 strategy aimed to discard any channels that were likely to be masked, leading to less overall stimulation and the potential for more spectral information at higher frequencies. Lai and Dillier (2008) compared the MP3000 map with ACE and found no benefit on an instrument identification test or an adaptive pitch ranking task using clarinet tones, using 2 CI users. The Harmonic Single Sideband Encoder (HSSE) strategy, which aims to encode TFS was used with 5 NHL using simulations and 8 CI users (Li *et al.*, 2013). They used the UW CAMP timbre and melody tests and whilst there was generally little improvement with the melody test, all CI users showed an improvement with timbre recognition. Temporal fine structure appears to be required to improve the appreciation of music, and whilst some TFS can be coded in these strategies

described above, the problem lies in how to successfully deliver this to CI users in a way that they can utilise (Li *et al.*, 2013).

3.3.2 Spectral channel availability

The best way to deliver frequency specific information to the implant is to use place pitch represented tonotopically as it is in the healthy cochlea, however not all active electrodes can be used independently at once. Available spectral channels can be limited by channel interactions. These channel interactions tend to occur more often with monopolar stimulation rather than bipolar or tripolar (Bierer, 2007), and simulated channel interaction has been shown to affect performance on the MCI in NHL (Crew, Galvin and Fu, 2015). Landsberger and Srinivasan (2009) compared tripolar and monopolar stimulation and found that a greater number of virtual channels could be created with tripolar. Koch *et al.* (2007) showed that an average of 7.1 different virtual channels could be heard between electrode pairs.

Two strategies designed to deliver high spectral resolution to the CI to maximise the virtual channels, the SpecRes and the SineEx were not shown to be advantageous over HiRes (Nogueira *et al.* 2009). They found no significant improvement using tests of speech intelligibility, an adaptive frequency difference limen test and speech and music appreciation questionnaires.

Current steering also has the potential to stimulate higher and lower than the most basal and apical electrodes, by distributing current to two intra cochlear and one extra cochlear electrode, to produce 'phantom electrodes'. Saoji and Litvak (2010) reported success with creating a percept that stimulated an apical area using partial bipolar stimulation, which sounded lower than what was achievable with monopolar stimulation.

3.3.3 Electroacoustic stimulation

Preserving acoustic hearing in suitable candidates in order to allow for EAS/bimodal stimulation means that better pitch cues are possible, as they allow access to high frequency sounds electrically and low frequency sounds acoustically, and these natural low frequency cues are able to deliver TFS and therefore provide essential pitch cues. Generally the addition of acoustic hearing with electric hearing leads to improvement. Significant benefits have been seen on melody recognition tasks when comparing 27 long electrode CI users with 5 short electrode hybrid users, using an open set melody recognition test (Gantz, Turner and Gfeller, 2006). Using a hearing aid in the contralateral ear (alone or as a combination with CI) has shown improvements on a familiar melody identification task with 9 CI users (Sucher and Mcdermott, 2009). Pitch ranking has also

shown great improvement when listeners use electro-acoustic stimulation (EAS) rather than CI alone, using the MedEl MuSIC Test, with a pitch ranking task using sine tones or piano, but not strings (Brockmeier *et al.,* 2010). However, Cullington and Zeng (2010) demonstrated no significant benefit of bimodal stimulation over bilateral implantation with the Montreal Battery for the Evaluation of Amusia subtests, a talker identification test or hearing in noise test when comparing 13 EAS with 13 bilateral CI users, however the tests used in this study may not have been sensitive enough to detect any differences.

3.4 How successful is pitch perception with CI?

Generally, pitch, timbre and melody perception are all very poor in CI users, whereas rhythm perception is often comparable to the performance of NHL (Gfeller and Lansing, 1992; Gfeller *et al.*, 1997; McDermott, 2004; Nimmons *et al.*, 2008) and audiologically matched (e.g. bilateral sensorineural hearing loss of 55 dB HL or worse across frequencies 1 – 4 kHz, and CUNY sentence score in quiet of <70% in best aided condition) hearing aid users (Looi *et al.*, 2008). Brockmeier *et al.* (2011) used the music the MedEl MuSIC Test and found no significant difference using the rhythm test between 31 CI users and 67 NHL. They did find a significant difference between these groups when using the melody recognition and instrument recognition tests. Nimmons *et al.* (2008) reported scores of average scores of around 50% for timbre recognition, with intra-family confusions for example when CI users heard woodwind they would typically choose brass or strings as their response. Cullington and Zeng (2010) tested 13 bilateral CI users and 13 bimodal users using the Montreal Battery for the Evaluation of Amusia (MBEA) and reported that pitch tests scored around chance and far below NHL, whereas both bilateral and bimodal users performed well on the rhythm test, almost as well as NHL. Similar results were reported in Cooper, Tobey and Loizou (2008) using 12 CI users.

One of the earliest reports of CI user pitch ability is that of Gfeller *et al.* (2002) who reported CI users pitch perception ability ranging from 1 to 24 semitones, with an average of 7.56 semitones. This was established using a pitch ranking task with synthesised piano tones ranging from73 Hz (D2) to 553 Hz (C#5), with a 2 AFC task. Participants were required to achieve at least 9 out of 11 trials correct before the interval was reduced. A score of 9 out of 11 when chance level is 50% (for a 2 AFC task) means that the likelihood of achieving that score by chance alone is 0.0327, which is less than the 0.05 level. The smallest interval tested was one semitone and the largest was 12 semitones, and so the range reported may not be representative of the CI users' abilities but rather of the limitations of the test.

More recently reported upper ranges for pitch ranking have been smaller, e.g. 6, 8 or 12 semitones (Drennan *et al.* 2010; Kang *et al.* 2009; Jung *et al.* 2010, respectively). Smallest intervals of lower than 1 semitone have also been reported for some CI users, however this was using a pitch discrimination task, rather than a ranking task (van Besouw and Grasmeder, 2011). Plus other studies have shown ceiling effects, with participants achieving scores of 1 semitone, however the tests were not able to test smaller than one semitone (e.g. Nimmons *et al.*, 2008; Jung *et al.*, 2009; Kang *et al.*, 2009; Maarefvand, Marozeau and Blamey, 2013; Drennan *et al.*, 2015). There are also individuals who are able to perform close to the abilities of NHL on some perceptual accuracy pitch tests, with abilities to discriminate intervals as small as one semitone or less (Gfeller *et al.*, 2002; van Besouw and Grasmeder, 2011). Average pitch ranking scores using UW CAMP have ranged from 2.4 – 8.1, with averages of 3.4 semitones (combined average of means reported in Nimmons *et al.*, 2008; Jung *et al.*, 2009; Kang *et al.*, 2009; Drennan *et al.*, 2015).

These abilities would indicate that for some CI users, the limited temporal information that is available through the CI is enhancing their listening ability, as being able to pitch discriminate or pitch rank for intervals of 1 semitone or smaller would not be possible by using place pitch cues alone.

Chapter 4 Measuring music perception

4.1 Why measure music perception?

Firstly, why music? Although CI were designed to provide a representation of speech, which is the most important reason for the CI intervention and the largest contributor to increase in quality of life, other non-speech sounds are an important part of life. Environmental sounds, such as the telephone, an ambulance siren, or a baby's cry are important to understand and respond appropriately to certain situations and an accurate representation of these can enhance the listening experience, if not be considered essential for quality of life and safety. It has been reported that music is considered the next most important thing after speech for CI users (Stainsby *et al.*, 1997). Although music testing may not be suitable for, or even desired by every CI user, there is a demand by both CI users and clinicians to be able to measure it. Secondly, why perception? Whilst electrodograms and vocoders can be used to visually and acoustically evaluate how well a CI transfers music, a test of perception takes into account the holistic approach: from the limitations of the CI processing, the hearing impaired ear, the interface between the device and the cochlea, and the acoustic and musical experience both before and after implantation. Finally, why measure? CI users are often subjected to speech testing: sentence, word and phoneme tests. Although much of the critique within this literature review could also be applied to speech testing, the history of CI research has seen the necessity to develop tests that are more difficult as CI users achieve more and more. A valid measure is therefore needed in order to establish the current levels of ability (Gfeller and Lansing, 1991, 1992) and to direct and assess any improvement.

4.2 How should music perception be measured?

There are a number of ways that information can be obtained about how well music is transferred to the listener via a CI. Self-report measures e.g. questionnaires are subjective, whereas imaging and event related potentials (ERP) are objective but may not be a true representation of perception of music. Psychoacoustic type tests of pitch perception aim to measure perceptual accuracy but the success of this depends on methods used, as well as the motivation of the listener. This section will present these methods and comment on their suitability for obtaining perceptual information with CI users for assessment and evaluation.

When using a holistic view of pitch perception, should music perception accuracy or music appreciation or satisfaction be the ideal outcome measure? Are they necessary for the other's success? CI users have reported enjoyment whilst listening to music, even with suspected or expected 'poor' perception (Lassaletta *et al.*, 2008). Measures of perception have not shown

strong relationships with measures of appraisal (Gfeller *et al.*, 2008; Bradley, 2010). However, 'when music perception is the criterion of implant benefit, accuracy of perception is only one criterion. The other criterion is the appraisal of the percept. That is, although the ability to recognize a musical excerpt can reflect an important outcome of implantation, whether the implant user evaluates the musical signal positively will determine whether they choose to listen to music and whether they evaluate the outcome of the implant favourably' *(*Gfeller *et al.*, 2008, *pp 9*).

4.2.1 Self-report questionnaires and appraisal

Questionnaires have focussed on music listening habits and appraisal before and after CI or hearing aid (HA) (Mirza *et al.*, 2003; Looi and She, 2010) and have included questions on perception, experience and education (Brockmeier *et al.*, 2010), before and after a change in strategy (Filipo *et al.*, 2008) and also to compare across strategies (Brockmeier *et al.*, 2007) and between NHL, bilateral and unilateral CI users (Veekmans *et al.*, 2009). Questionnaire based research can be biased by the motivation of the individual, internal criterion and expectations and this can lead to difficulty quantifying any significant improvement. Measures of enjoyment that ask CI users to appraise certain musical pieces are inherently problematic because of previous listening experience, musical likes and dislikes, and are discussed by Mehra, (2012) as personal, situational, cultural and emotional factors. Large variations exist in musical appraisal (Gfeller *et al.*, 2000; Galvin, Fu and Nogaki, 2007) and thus it is difficult to generalise about the CI population.

4.2.2 Imaging and event related potentials

By using objective recording techniques during listening tasks, specific areas can be associated with perception and indications can be gained regarding perceptual discriminations. This can provide information about the limits of what the auditory system is able to hear in terms of absolute sensitivity or sensitivity to change. Koelsch *et al.* (2004) used ERPs to determine if certain irregularities in music could be detected by CI users. They tested 12 CI users and 12 NHL and presented chord sequences that occasionally contained chord patterns that were irregular, which usually evoked certain early and late cortical responses (the Early Right Anterior Negativity (ERAN) and the N5 peak). They demonstrated that these occurred for both NHL and CI users, although the amplitudes were smaller for CI users.

Limb *et al.* (2010) presented melody, rhythm and language tasks to 10 CI users and 10 NHL. Comparisons were made between rest and task for both groups, and also between CI users and NHL.

Results indicated that temporal regions seemed to be highly activated for the CI users for each of the tasks, with larger activation than was seen with NHL. This study provided insight as to areas that may be involved (by correlation) with music and language tasks, and highlighted the potential of thos method to be used to determine limits of perception for CI users. It was stated by the authors that neuro-imaging was not easy with CI users given the nature of electromagnetic limitations that were associated with implantation and also the generation of noise levels which can interfere with testing, unfortunately making it problematic for determination of ability or validation of other methods.

These studies highlight alternative and objective methods by which the music perception of CI users can be investigated. An ERP or imaging study provides a method which may compliment other methods of investigating music perception, however, some of the findings are similar to that of discrimination (rather than ranking) tasks: a response may only determine the ability to detect a change. It is not clear whether these methods can be developed into ways to provide more detailed information about more holistic aspects of music perception. Musical and language experience can also affect the responses to such testing, meaning that other aspects may influence results and differences may exist at baseline (Jeng *et al.*, 2011).

4.2.3 Psychoacoustic perceptual accuracy methods

A well designed measure of perceptual accuracy should be able to provide a comparable result across different people about the range of perceptual abilities in a population. Historically, psychoacoustic tests have provided much information regarding the perceptual ability of CI users. More recently, a number of methods have been employed to assess the potential for CI strategies to transmit information essential for music perception. Schroeder phase discrimination stimuli, which can be determined by differences in TFS alone, can be used in CI users to evaluate the delivery of TFS by the CI device. Drennan *et al.* (2008) showed that CI users are able to use timing differences across channels in order to discriminate the Schroeder phase stimuli. Won *et al.* (2010) showed that Schroeder phase stimuli can be used to distinguish between two processing strategies that are known to provide differences in TFS transmission, thus supporting the use of this test in strategy evaluation and suggesting sensitivity to changes of this magnitude. However, Schroeder phase discrimination stimuli tests have shown no correlations with a music perception test designed for CI users: the UW CAMP test (described later), which may be related to the lack of rhythm section in this test battery (Won *et al.*, 2010). This suggests that either the Schroeder phase stimuli discrimination test is measuring a separate mechanism to what is being measured by the UW CAMP, or that the UW CAMP is not as sensitive to the changes which may be small.

The spectral ripple test measures ability to resolve differences in frequency. Participants are required to differentiate between different ‘ripples’ thus determining threshold for spectral ripple

discrimination. Results with 31 CI users have shown significant correlations with speech in noise tests, which adds weight to the validity of the test in terms of frequency resolution. Test retest reliability was shown to be high, resulting in this test being suggested by Won *et al.* (2010) as a useful way to evaluate between processing strategies and that can potentially be conducted in the clinic. Further work in this area showed the spectral ripple test to correlate with all three aspects of the UW CAMP (Won *et al.*, 2010). The spectral ripple test appears sensitive enough to highlight differences in strategies and shows correlations with music perception tests, but it is not clear as to whether this test could be used other than to determine differences in performance. Results from the spectral ripple test may not be able to guide clinicians regarding tuning or provide any understandable results to feed back to a CI users.

In summary, these methods are not considered ideal for the accurate measure of pitch perception in CI users. Self-report methods are too subjective and subject to bias, and the equipment, time and training needed for both imaging, ERP and psychoacoustic studies means that these will not be ideal for use in the clinical environment. Results may not be easily applicable to musical related problems, and explaining results to CI users may not be straightforward, nor may the results clearly guide a clinician in changes to map or strategy.

4.3 Music perception tests

The last 30 years have brought about the possibility of testing the music perception of CI users which has allowed clinicians and researchers the ability to record and monitor progress of music perception. More recently, manufacturers cite new designs for CI hardware and software that aim and claim to improve music perception. Suitable and reliable music perception tests are essential for the assessment of these developments. Many music perception test batteries consist of a number of aspects considered important to music perception, often including tests of pitch, rhythm and melody perception. This thesis is focused on the perception of pitch discrimination and pitch ranking, and so test batteries that include those tests will be described here, in chronological order.

4.3.1 The Primary Measures of Music Audiation (PMMA)

The PMMA was developed by Gordon (1979) to assess the musical potential of school children from kindergarten (age 5) to grade 3 (age 8). It was designed to test 'audiation' which is the ability to 'mentally' hear and comprehend music, when the actual sound is not present. Comparisons were

made between 873 children from a standardised school, 77 children from a community music school and 75 children from a private academic school. Significant differences were seen between the standard school and the private school for the tonal test, leading Gordon to suggest that audiation depends both on innate ability and environmental cultural influences, and as a result, he states that this indirectly validates the PMMA. Gordon reports reliability on retest as ranging from r = 0.70 – 0.91, although a later study using Greek children (n = 1188) reported much poorer retest scores, ranging from r = -0.1 – 0.64 (Stamou, Schmidt and Humphreys, 2010).

The PMMA was chosen to be included in this thesis because it was one of the earliest test of pitch perception to be used with CI users (Gfeller and Lansing, 1991, 1992) and has been used with CI users on several other occasions since then (Lassaletta *et al.*, 2008; Edwards, 2013). It consists of a tonal and rhythm test, with both tests presenting 40 pairs of note sequences ranging in pitch from C4 (261.6 Hz) to C5 (523.3 Hz) and the listener is asked to determine whether each pair was the same or different.

Gfeller and Lansing (1991, 1992) used the PMMA for use with CI users and this was the first documented test of music perception to be used with CI users. Average scores for 34 CI users were 77.5% for the tonal test, with no reported floor or ceiling effects being seen. No significant differences were seen between devices of the time: Ineraid: 75.6 %, Nucleus: 79.5 %. It is not clear whether this indicated that no pitch perception differences existed between the devices, or whether the PMMA was not sensitive to these changes. Content reliability in terms of split halves measures was shown to be high (0.79) and the Kuder-Richardson Formula 20 (KR-20, Kuder and Richardson, 1937) which measures internal consistency was also high (0.68), although CI users were not tested twice in order to establish test-retest reliability for CI users. Gfeller concluded that with minor modifications relating to items and visual prompt, the PMMA 'was a usable measure for quantifying melodic and rhythmic discrimination'. Similar results were seen in a later study by Lassaletta *et al.*, (2008) who found mean results of 71% for the tonal test and found no association with self-reported musical enjoyment in a study using 65 adult CI users. Edwards (2013) used the PMMA with HI children and compared bilateral CI with bilateral HA and bimodal stimulation, and found mean scores of 58% (23/40) for the CI children on the tonal test, which was significantly worse than the other groups. The Gfeller and Lansing (1991) study reported ranges of 65-90%, and although no ranges were reported in the Lassaletta *et al.* study, the standard deviation (SD) was 13.6%, indicating that within one SD, ceiling effects were not occurring, and that the level of difficulty of the PMMA was suitable for CI users. No studies have reported test-retest reliability with CI users, or any evidence of validity.

Based on the PMMA alone, the average pitch ability of CI users would be estimated to be 71% – 77.5% for adults (Gfeller and Lansing, 1992; Lassaletta *et al.*, 2008) and 58% for children (Edwards, 2013). There is very little information about the PMMA regarding the level of difficulty within its 40 questions and so without further investigation it is not easy to put these results into context, other than making comparisons with normal values. The only validation attempts with CI users relate to good split halves reliability and internal consistency, and some evidence to indicate that floor and ceiling effects were not present (Gfeller and Lansing 1992).

4.3.2 Minimum Auditory Capacity Arena (MACarena)

The MACarena is an auditory testing software platform, originally developed by Lai and Dillier (2002) to be a flexible computer based speech testing platform for use with hearing aid and CI users. It was adapted by Buechler, Lai and Dillier (2004) specifically to assess the music perception abilities of users of bimodal stimulation. There appear to be no further reports of use of the MACarena pitch test in the literature. The MACarena included tests of pitch, rhythm, interval and contour discrimination, song recognition and instrument and musical quality. The pitch test presented 24 pairs of notes, and participants were asked to decide if these pairs were the same or different. Sixteen of the pairs were different, and 8 were the same. Of these 16 different pairs, 14 differed by 1 semitone and 2 differed by 2 semitones. Having 14 repeats for the interval 1 semitone was good, because the greater number of repeats, the more statistical confidence can be placed in the results, however having only 2 repeats for an interval of 2 semitones did not provide sufficient statistical confidence.

Buechler, Lai and Dillier's (2004) MACarena results with 10 bimodal stimulation candidates showed that with unilateral CI, results ranged from 25 to 80% correct, of the 24 pairs of 1-2 semitone intervals, with an average of 60%. Chance levels were 50%, thus these CI users were getting an average score of just above chance. Scores for bimodal stimulation (e.g. CI in one ear, HA in the other ear) were similar to HA only, with average scores of 75% in both conditions. These scores are easier to interpret if the stimuli difficulty is known (which it is); e.g. these participants scored averages of 60-75% with a task that was almost entirely testing one semitone, and did so with 14 repeats. No ceiling effects were seen for this group, unlike for the NHL, and because some participants scored as low as 25%, floor effects did appear to affect this group. No further validation or test retest reliability is described. As no statistical analysis was conducted between the conditions, no comment can be made on whether the test is successful at differentiating between such groups in a significantly statistically significant sense. The authors reported that as part of a wider test battery within the MACarena testing platform, it appeared to be a valuable resource for the testing of music perception for CI and

HA users. The benefits of the MACarena pitch test are the use of the MCS, the relatively large number of repeats (14) for the interval 1 semitone. However, use of the stimuli from Schuppert *et al.* (2000) may well exclude a large proportion of CI users who may be unable to successfully differentiate between intervals of 1 or 2 semitones.

4.3.3 Montreal Battery Evaluation of Amusia (MBEA)

The MBEA test battery (Peretz, Champod and Hyde, 2003) was designed specifically to evaluate the musical abilities of people with brain damage, as some brain anomalies or injuries will not always affect musical skills even when individuals have suffered cognitive loss. It consisted of six tests, of contour, interval, scale, rhythm, meter and memory, all set within short phrases of music, 'with sufficient complexity to guarantee its processing as a meaningful structure rather than as a simple sequence of notes' (Peretz, Champod and Hyde 2003, p62). It was included within this thesis due to its inclusion of a pitch based test, and its repeated use with CI users (Hopyan *et al.*, 2012; Cooper, Tobey and Loizou, 2008). The first three tests all look at a pitch based component of testing: the 'scale' test modified one note within the melody so that it was out of scale, although still within the same contour; the 'contour violation' test changed one note so that this note was creating the opposite contour, however was still in the same key and the final test was the 'interval violation' test, which altered the interval but the different note was still in the same contour and in the correct key. All three pitch based violations had similar average interval changes, of 4.3, 4.3 and 4.2 semitones. Due to time constraints only one subtest could be chosen from the MBEA and the scale test was chosen as it was felt to be the simplest violation and the closest one to other tests of pitch discrimination.

The scale test presented the listener with two short musical phrases of 7-21 synthetic piano notes, ranging from 247 Hz (B3) – 988 Hz (B5), with an average of 5.1 seconds long. Within that phrase one note was different in 15 of the 31 phrases, 15/31 were identical and 1/31 phrase was a deliberate odd phrase and was not included within the scoring. The listener had to decide whether the phrases were the same or different. Analysis of the MBEA regarding the magnitude of interval change between phrases 1 and 2 was complicated. The difference between the changed notes from phrase 1 to phrase 2 ranged from 4 to 6 semitones, however a substantial amount of time had passed between the presentations of these 2 notes when compared to other more simple tests. Essentially the task asked the listener to compare 2 notes that were hidden amongst several other notes.

Sixty-eight neurologically normal listeners were tested with the MBEA, and the average score for the scale subtest was 27/30. Test-retest reliability, using a composite score which combined all 6 subtests, was also conducted using 28 neurologically normal listeners 4 months after initial testing.

Whilst these listeners tended to improve at time 2, a significant correlation of $r = 0.75$ between time 1 and time 2 supported test-retest reliability. Peretz *et al.* (2003) report that the test has high-sensitivity because 'more than 80% of the participants do not obtain a perfect score' (*p*65) however this sounds more related to low levels of ceiling effects, indicating that difficulty is appropriate for this group of listeners. The definition for 'amusia' was two standard deviations below the normative values, and this was determined as 78% by Peretz *et al* (2003).

The MBEA has also been used alongside neuroimaging techniques (Mandell, Schulze and Schlaug, 2007), using 51 healthy listeners. Average scores for the scale test were 82%, but the high correlations between the MBEA and the neuroimaging techniques offer insight into the neurophysiological basis for amusia, and therefore also offer validation for the MBEA as a test of amusia. Cuddy *et al.* (2005) also used the MBEA with self-reported amusics, however found that there wasn't a significant difference between scores for these listeners when compared to non-amusics, using the scale test.

The possibility of using the MBEA with CI users was investigated by Cooper *et al.* (2008). They used 12 CI users and 30 NHL using CI simulations and found that CI users typically performed at chance level for the scale test (median score of 17/30) and they state that modifications may need to be made due to the difficulty CI users have with the pitch tests of the MBEA. As part of their assessment for the suitability of the MBEA for use of CI users, Cooper, Tobey and Loizou (2008) did not report on test-retest reliability and although no validity was expressly discussed, comparisons with the CI simulations indicated that CI users were performing at a similar level to the NHL who were using 4 to 6 independent channels.

The MBEA was also used in a bimodal study (Cullington and Zeng, 2010) as part of a battery of tests comparing the performance of bimodal with bilateral CI users. Mean scores for the scale test were just above chance (17.5/30, e.g. 58.3%) for the bilateral CI users. Whilst performance across the MBEA was better with bimodal, results were not statistically significant. It is unclear whether this indicates that the MBEA is not sensitive enough to detect such a difference, which may be subtle. It is interesting to note that other clinical measures of speech perception showed no significant difference either, which may suggest that no or minimal differences existed in this sample. Alternatively, the measures used may be too difficult for the population(s) being tested. The child's version of the MBEA has also been used, Hopyan *et al.* (2012) used 23 CI children and found a median score of 11/20 for the scale test.

Based on these studies, CI users, unilateral, bilateral, bimodal and child CI users were all performing at around chance level for the scale test, which most likely reflects the difficulty and

the subtlety of the differences in the MBEA scale test. This indicates that in its current form it is not suitable in terms of complexity and difficulty for CI users: not only are the intervals very small, but successfully discriminating them is made more difficult by the short-term memory required to succeed on the task. The use of piano tones and the relatively low frequency range may also be adding to the complexity experienced (Cooper, Tobey and Loizou, 2008). No publications evidence validity or reliability for use with CI users.

4.3.4 MedEl Musical Sounds in Cochlear Implants (MuSIC) Test

Fitzgerald and Fitzgerald (2006) designed a test battery that included tests of pitch identity ranking, rhythm, melody, distinguishing chords, instrument identification, emotion and dissonance, using recordings of a choice of real instruments.

A small subset of these tests was used by Arnoldner *et al.* (2007) to assess musical benefits of upgrading from CIS to FSP, however they did not use the pitch identity ranking test. Brockemier *et al.* (2010) used all 8 subtests with 13 EAS, 13 unilateral CI and 13 NHL. The pitch identity ranking test employed three interweaved tests: piano with target note of 262 Hz (C4), strings with target note of 440 Hz (A4) and pure tone with target note of 349 Hz (F4). The EAS participants were tested in EAS and CI only conditions and no significant differences were seen in scores. Scores were significantly better with EAS compared to CI on piano and pure tone pitch ranking. Unsurprisingly, the NHL were significantly better than the CI users on the instrument detection and identification. Hours listening to music prior to hearing loss for EAS was significantly positively correlated with the pitch identity ranking scores for all three instruments. The hierarchy seen between NHL, EAS and CI users indicated some criterion validity, e.g. that the test was performing in a way expected due to the differences in ability of these three populations. Test retest reliability does not appear to have been addressed in this study, and large variations may account for test differences seen that do not provide statistical significance.

Further evaluation (Brockmeier *et al.*, 2011) was conducted using 31 MedEl CI users and 67 NHL. The pitch identity ranking test was used with piano tones (target note of 262 Hz, C4), strings (target note of 440 Hz, A4) and pure tone (target note of 349 Hz F4). Significant differences were seen between average NHL and CI users' scores for pitch identity ranking with all 3 timbres. Piano scores were significantly poorer than pure tone for CI users, and significantly poorer than both other timbres for NHL. Median scores for CI users for piano were 10.3 semitones, strings 8.4 semitones, and pure tone 5.7 semitones. The lower ranges of these scores were at 0.5 semitones for all 3 timbres, indicating ceiling effects in this group, and upper ranges of 17.5 semitones for piano, 15 semitones for strings and 14.5 semitones for pure tones. As the MedEl MuSIC Test included much wider intervals than those intervals, floor effects were not seen in this group.

Interestingly, the ranges for NHL were also quite wide: 0.5 to 11 semitones for piano, and 0.5 to 4 semitones for strings and pure tones. This is discussed in Brockmeier *et al.* (2011) as potentially being due to difficulties with non-musician NHL and pitch ranking tasks, and they refer to the approximate 4% of NHL with amusia. Test-retest reliability was assessed using 9 NHL and 9 CI users over a period of 6 weeks – 3 months, and no significant differences were seen on retest, however ceiling effects may have influenced these results.

4.3.5 The Melodic Contour Identification (MCI) test

Developed by Galvin, Fu and Nogaki (2007) this test aims to measure pitch identity perception within the context of a melodic contour. Utilising 9 different melodic contour patterns (Figure 6.1), participants must chose the contour they heard.

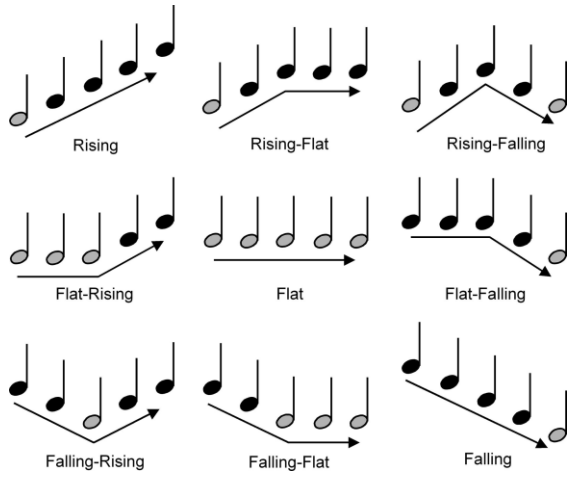


Figure 4.1 The Melodic Contour Identification test contours, from Galvin, Fu and Nogaki (2007), reproduced with permission

Stimuli used was a three harmonic complex tone, with three base notes A3 (220 Hz), A4 (440 Hz) and A5 (880 Hz). Interval gaps of 1, 2, 3, 4 and 5 semitones were used, with mean scores for the 11 CI users of 32%, 52%, 58%, 61% and 64% for each interval. One semitone was significantly poorer than the other intervals. Mean results were also compared across base notes, and A3 was found to be significantly poorer for CI users. They also found that MCI training, using a different note range, conducted over 1 week – 3 months significantly improved MCI scores in 6 of the original 11 CI users. This increased ability was maintained at one month follow up post training. The positive effect of training has also been shown by Lo (2014), who developed a melodic contour training program and showed improvement post training with the 16 CI users being better able to identify

melodic contours with shorter note durations and smaller interval sizes. Similar results were shown by Cheng *et al.* (2018), using 16 Mandarin speaking children CI users: MCI scores improved with 8 weeks of training and were still maintained at 4 weeks post training.

The role of timbre has been investigated using the MCI: Galvin, Fu and Oba, (2008) compared the organ, glockenspiel, trumpet, clarinet, violin and piano and found that MCI scores with the organ were better than piano for both 8 CI users and 8 NHL, although neither to the point of significance. A significant effect of timbre on MCI score was seen for 5/8 CI users when analysed individually. Scores for CI users and NHL were poorer than with the 3 tone complex version of the MCI used in Galvin, Fu and Nogaki, (2007). In addition, when the MCI was presented with any of the 6 instruments, better scores were not seen with the largest intervals; indicating non-monotonic psychometric functions, and suggesting that the extra instrument harmonics were causing confusion. The addition of a competing instrument masker (as a flat contour) played simultaneously to the MCI task resulted in much poorer scores for CI users but not NHL (Galvin, Fu and Oba, 2009). Timbre of MCI stimuli can also have a detrimental effect on non-musician NHL: Crew, Galvin and Fu (2015) developed the sung speech corpus (SSC) which was based upon the MCI, and used 50 mono-syllabic words, making up 5 word sentences and sung to the contours of the MCI, using intervals of 1, 2 and 3 semitones. They conducted their study using 16 NHL, 8 musicians and 8 non-musicians, and whilst musicians excelled at the tasks, non-musicians had much poorer average scores with the sung speech when compared with the piano tone used in Galvin, Fu and Oba (2008)

The MCI has also been used in a study that coupled it with the mismatch negativity auditory evoked potential (MMN AEP), testing 10 CI users and 10 NHL with a variation of the MCI using 1 or 5 semitone intervals (Zhang, Benson and Cahn, 2013). An oddball procedure was created whereby the falling (or rising) contour was presented 340 times and amongst these falling or rising contours, a different contour (e.g. the falling flat or rising flat) was presented 60 times, creating the MMN AEP. In NHL, when the falling flat or rising flat contour was heard (e.g. the oddball), the MMN was recorded for 7/10 NHL in the 1 semitone condition, and for 8/10 NHL for the 5 semitone condition, for both the falling/falling flat and rising/rising flat conditions. In contrast, no MMNs were recorded for any of the CI users for the 1 semitone condition, whereas 3/10 showed an MMN for the 5 semitone falling/falling flat condition, and 6/10 CI users showed an MMN for the 5 semitone rising/rising flat condition.

The MCI has also been adapted to create contours using loudness, and loudness and pitch contours combined (Luo, Masterson and Wu, 2014). They tested 10 CI users, and 6 NHL and found that NHL were less affected by whether the contours were created by pitch, loudness or a combination of

pitch and loudness, whereas CI users performed better when the pitch and loudness was combined. This was significantly better for the 1 semitone condition, indicating that when pitch and loudness cues are in agreement (e.g. pitch goes up and loudness goes up), CI users can take advantage of these extra cues especially when pitch cues are weak and the task is difficult.

4.3.6 University of Washington Clinical Assessment of Music Perception (UW CAMP)

The UW CAMP was designed by Nimmons *et al.*, (2008) to test the music perception of CI users. It includes a pitch ranking test, a melody recognition test and a timbre test. The pitch ranking test presents two tones in a 2 AFC task, and the listener has to decide which tone is higher in pitch. Piano-like tones, of 760 ms, were created for the test, and were designed to have identical spectral envelopes, created from a piano note of C4 (262 Hz), and uniform temporal envelopes, in order to minimise temporal envelope cues. The test used an adaptive 1-up 1-down procedure which tracked the psychometric function at 50%. Four base notes were interleaved: F#3 (185 Hz), C4 (262 Hz), E4 (330 Hz) and G4 (392 Hz). Each trial started by comparing the base note to the note 12 semitones above it, and if the listener chose the correct answer, the interval was reduced to 6 semitones, and then to 1 semitone. If they got the answer wrong, the comparison interval was immediately reduced to 1 semitone. If the listener successfully pitch ranked at 1 semitone, the algorithm added an automatic reversal at '0' semitones, in order to satisfy the need to have enough reversals to terminate the run. This occurred after 8 reversals, and the result was calculated as the mean of the last 6. This was repeated 3 times for each base note, and the final answer was the average of those 3.

Four different base notes were chosen because of the possibility of frequency affecting pitch perception in CI users, and notes surrounding the octave of middle C (262 Hz) were chosen because this is a common octave for both western musical instruments and traditional nursery rhymes. The UW CAMP was not designed to test intervals smaller than 1 semitone because it was the smallest interval in typical Western music scales. Nimmons *et al.*, (2008) used 8 CI users, and the results from the different frequencies can be seen in Table 4.1.

The first validation of the UW CAMP was undertaken by Kang *et al.*, (2009). They established test retest reliability using 35 CI users and reported an Intraclass Correlation Coefficient (ICC, unspecified and undefined) of 0.85 for the pitch test. It is not clear from the paper whether this is a combined ICC across all frequencies tested. They demonstrated moderate correlation between pitch test results and speech reception thresholds (SRT) in noise, and moderate to strong correlation between pitch test results and CNC, thus they claimed the UW CAMP showed construct

validity, and they also stated that the significant difference seen between CI users and NHL demonstrated concurrent validity. It is argued by this researcher that a pitch perception test would have to be very poor to not be able to distinguish between such groups. Further correlations between UW CAMP pitch test results and speech performance scores were shown by Jung *et al.* (2009), using 12 Korean speaking adult CI users. Floor and ceiling results were seen in their results, with scores of < 1 and 11.54 semitones being reported.

Drennan *et al.* (2008) used Schroeder phase stimuli, which are 2 sets of stimuli with identical spectra and minimal envelope modulations and differ only in their TFS, and so if CI users are able to discriminate between them it would suggest that they are hearing and utilising TFS. Their aim was to relate these findings to clinical tests that may benefit from TFS e.g. the UW CAMP. They tested 22 CI users and reported that the results from base note E4 (330 Hz) correlated with the 200 Hz Schroeder phase discrimination stimuli results, and that the C4 (262 Hz) and the average of all 4 notes of the UW CAMP correlated with the 400 Hz Schroeder phase discrimination stimuli.

Won *et al.* (2010) also showed relationships between UW CAMP pitch scores and spectral ripple discrimination, with CI users who had better spectral ripple discrimination performing better on pitch tests. Won *et al.* (2010) used a different method to Nimmons *et al.* to calculate the final result from the UW CAMP pitch test; they used the Spearman Karber method to determine the 75% tracking level on the psychometric function, rather than the 50% tracking point, but showed these 2 sets of results to be highly correlated. Similar correlations between spectral ripple and UW CAMP were not shown in a group of 11 paediatric CI users (Jung *et al.*, 2012).

Wright and Uchanski (2012) compared the MBEA, the MCI and UW CAMP using 10 CI and 27 NHL using CI simulations, and found no correlations between the subtests of these music perception tests, highlighting the importance of information regarding how dissimilar these tests are and the lack of comparability between tests methodologies and test results. They concluded that the tests are assessing very different abilities, and that they tax the listener in different ways. They state the importance of careful consideration regarding test choice, depending on the research question.

Golub *et al.* (2012) used the UW CAMP to compare the performance of 5 users of hybrid devices. They compared their results with previously published results of CI users, in addition to comparing full hybrid with acoustic only results. They found hybrid users were significantly better at the UW CAMP pitch test than CI users at 262 Hz, they also found at this frequency that acoustic stimulation alone was better than hybrid results.

Finally, a large study conducted by Drennan *et al.* (2015) using 145 CI users showed that music training had no effect on the UW CAMP pitch test. They provide validation to previous UW CAMP

test results, as very similar results to the previous papers were reported with a much larger cohort. They argue that the use of the 1 up 1 down adaptive procedure that tracks at the 50% point on the psychometric function is suitable because it was correlated highly (*r* = *0*.91) with the Spearman-Karber method to compare to the 75% point on the psychometric function. They concluded that because these two estimation points were highly correlated, and because Kang *et al.* (2009) has shown an ICC of 0.85, that both methods were equally valid and suitable for use with CI users. They state this was done in reference to criticism of the use of the 50% tracking point (Maarefvand, Marozeau and Blamey, 2013), however this paper seems to be making more of a reference to the problem of non-monotonicity rather than the tracking point. In addition, Drennan *et al.* (2015) state that the significant correlations with CNC word recognition and SRT scores indicate the UW CAMP's evidence of reliable and clinically meaningful results. They also provide evidence of construct validity for the UW CAMP. They report its use in an a trial of the Nucleus Hybrid L (a hybrid cochlear implant model), showing that UW CAMP results with this hybrid are similar to NHL scores, whereas when no acoustic information was available, scores were much closer to average CI user performance. They also used the Kang *et al.* (2009) reliability data with their current data to construct 95% confidence intervals for the UW CAMP, and reported that a difference of 2.4 semitones would be considered a significant difference.

Table 4.1 Summarising the average scores for the UW CAMP in the literature

Study	Base note (Hz)	Mean (semitones)	SD (semitones)	95% confidence interval	Range (semitones)
Nimmons *et al.*, (2008) n = 8	185 Hz	2.9	2.5	Not reported	1-9.1
	262 Hz	3.8	3.7	Not reported	1-11.5
	330 Hz	3.3	2.4	Not reported	1-9
	392 Hz	2.4	2.4	Not reported	1-6.5
Kang *et al.*, (2009) n = 42	262 Hz	2.9	2.7	Not reported	Not reported
	330 Hz	3.4	3.1	Not reported	Not reported
	392 Hz	2.5	2.5	Not reported	Not reported
Jung *et al.*, (2009) n = 12	262 Hz	2.7	1.7	Not reported	0.8-6.9
	330 Hz	4.4	4.2	Not reported	0.65-12
	392 Hz	8.1	3	Not reported	1.8-11.5
Won *et al.*, (2010) n = 42 used Spearman Karber method, e.g. 75% tracking		5.1		1.6	Not reported
		5.0		1.6	Not reported
		3.7		1.7	Not reported
Jung *et al.,* (2012) n= 11 (paediatric)		2.9		1.6	Not reported
		3.0		3.0	Not reported
		3.1		2.6	Not reported
Drennan *et al.,* (2015) n=140		3.2		Not reported	Not reported
		2.6		Not reported	Not reported
		3.1		Not reported	Not reported

4.3.7 South of England Cochlear Implant Centre Music Test Battery

The South of England Cochlear Implant Centre Music Test Battery Pitch Discrimination Test (SOECIC MTB PDT) was developed by van Besouw (2010) at the SOECIC/Institute of Sound and Vibration Research (ISVR) for assessment and rehabilitation of CI users. Originally consisting of a pitch discrimination test with an identity component, and a rhythm discrimination test, both tests are presented to the listener using a three alternative forced choice procedure and an adaptive method to determine discrimination threshold. The PDT is able to measure up to 16 semitones

and down to 1 cent (= 1/100th semitone), with the resolution improving as smaller intervals are approached.

Although these small pitch intervals seem unnecessarily small for testing CI users, e.g. 1 cent, the limits of ability of CI users of 1 – 24 semitones that are reported (Gfeller *et al.*, 2002) were obtained using a measure that did not test smaller than 1 semitone. Studies using the SOECIC MTB PDT have also shown CI users' ability to achieve scores of lower than 1 semitone (van Besouw and Grasmeder, 2011). It is hoped that future CI technological developments are going to lead to improvements in music perception and if today's CI users are already achieving interval discrimination of less than 1 semitone, it is important that this level of absolute sensitivity is upheld.

Potential loudness cues during pitch testing are eliminated by roving the stimulus by ±3dB, although studies with NHL have said that this can be confusing when close to threshold (Lamb, 2010). Reliability, tested using 18 NHL, was shown to be better when using a 3 interval 3 alternative forced choice procedure (3I3AFC), when compared to a 3 interval 2 alternative forced choice procedure (3I2AFC) (Lamb, 2010). The test battery has also demonstrated significant differences between musicians and non-musicians in NHL (Paynter, 2010).

A study of nine adult CI users was conducted with the SOECIC music test battery to evaluate any change in results from upgrade from CIS to FSP strategies (van Besouw and Grasmeder, 2011). The FSP strategy is designed to provide more fine temporal information which should improve music perception. Testing using the music perception tests as well as BKB sentence tests in quiet and noise were conducted before upgrade and at 6 weeks subsequent. No significant differences were found between the processing strategies for music or speech tests. This lack of significant effect may be caused by the small sample size and large variability in CI users' scores, resulting in the test lacking in power to detect any change that may exist between these two groups.

This test benefits from having an extremely large range of abilities that it can in theory cater for, indicating that it can be used with both NHL and CI users which provides benefits of easy access to normative data without test modification. The test provides options for both pure and complex tones, as well as a choice of frequency range. There is no evidence provided in the literature to indicate test-retest reliability when using CI users. Although SOECIC MTB PDT scores have been correlated with BKB sentence tests, no attempt has been made to validate the test results. It utilises an adaptive measure, based on the transformed up down method (Levitt, 1971) and tracks the 71% level using a 2 down 1 up procedure. This type of method has been criticized previously

by Swanson (2008) because of the necessity of assuming the underlying psychometric curve is monotonic (Levitt, 1971).

4.4 Ideal test qualities

This section will describe the ideal qualities for a pitch perception test for cochlear implant users. The overriding goal was to establish whether existing and future tests are valid, defined as purporting to do what they are designed to do. However, under this umbrella definition are several types of specific validity including:

Content validity	Is the *content* of the measure appropriate?
Construct validity	Has the measure been *constructed or designed* in a suitable manner?
Criterion validity	Do the results *relate to* an established test (concurrent validity), or can they be used to *predict* performance (predictive validity)?
Face validity	Does it *appear* to test what it purports to test?
External validity	Are the results of the test *generalisable* to other populations?

Furthermore, other issues such as reliability also affect validity, and a measure cannot be considered valid if it is unreliable. Fitzpatrick *et al.* (1998) describe a number of important aspects for consideration when developing outcome measures.

Validity	Does it measure what it claims to measure?
Reliability	Are the results reproducible, and does the measure demonstrate internal consistency?
Appropriateness	Is the content appropriate?
Responsiveness	Does it measure changes that matter to patients?
Precision	Is the measure suitably precise?
Interpretability	Are the results easy to interpret?

Acceptableness	Is the measure acceptable to patients?
Feasibility	Is the measure feasible to administer and process?

External validity was disregarded as it was not considered necessary to generalise pitch perception results from CI users to other populations, and face validity was not considered important in this context as most pitch perception tests for CI users appeared to be face valid. It was therefore considered that 3 types of content validity were important: the type of pitch test, an appropriate level of difficulty and an appropriate stimuli choice and that 2 types of construct validity were important: a suitable test methodology for CI users, and a suitable number of repeats in order to achieve statistical confidence in the result. Reliability on retest was considered essential. Responsiveness and precision were considered to be within the realms of content validity: in terms of stimuli choice and difficulty level.

4.4.1 Content validity: type of pitch test

Tests of pitch perception can be further divided into tests of pitch discrimination, where listeners must merely distinguish between two sounds occurring one after the other, and tests of pitch ranking, where two sounds must be ranked into an order based on their tone height. An ideal pitch test for CI users would combine both of these.

4.4.2 Content validity: difficulty

An ideal test needs to have appropriate difficulty. Ceiling and floor effects can mean that the ability of a population cannot be accurately assessed. In addition to avoiding ceiling and floor effects, the sensitivity between the largest pitch and the smallest pitch that a test can assess, is also important, in terms of the number of intervals and the gap between intervals tested.

Gfeller *et al.* (2002) reported CI users' pitch perception ability ranging from 1 to 24 semitones, with an average of 7.56 semitones. This was established using a pitch ranking task with synthesised piano tones ranging from73 Hz (D2) to 553 Hz (C#5), with a 2 AFC task. Participants were required to achieve at least 9 out of 11 trials correct before the interval was reduced, resulting in statistical significance at the $p < 0.05$ level. The smallest interval tested was one semitone and so the range reported may not be representative of the CI users' abilities but rather of the limitations of the test. More recently reported upper ranges for pitch ranking have been lower, e.g. 6, 8 or 12 semitones (Drennan *et al.* 2010; Kang *et al.* 2009; Jung *et al.* 2010, respectively). Smallest intervals

of lower than 1 semitone have also been reported for some CI users (Gfeller *et al.*, 2002; Nimmons *et al.*, 2008; van Besouw and Grasmeder, 2011; Maarefvand, Marozeau and Blamey, 2013)

4.4.3 Content validity: stimuli choice

Due to the complexities associated with the CI interface with the hearing impaired cochlea, there are many factors that influence the success of the CI in terms of pitch perception. As a result, success with pitch perception is unlikely to be uniform across the electrode array and associated areas of the cochlea; and so tests that use different frequencies will provide more information. In addition, different timbres are also useful to establish the role of harmonics in terms of aiding or hindering pitch perception, and also different instrument timbres are likely to be of interest to the listeners themselves, depending on listening or playing preference. Whilst sine tones (consisting of the fundamental frequency only) are of interest, many tests have used complex tones only, as pure tones are seldom heard in the real world (Gfeller *et al.*, 2002; Galvin, Fu and Nogaki, 2007; Nimmons *et al.*, 2008).

4.4.4 Construct validity: suitable methodology

Methodologies used in existing tests are variations on either the method of constant stimuli, or an adaptive procedure. Benefits to an adaptive procedure over the method of constant stimuli are that they are more efficient: they achieve a threshold in much less time. Adaptive procedures used in psychometric situations are subject to certain assumptions. As described in Levitt (1971), in order to use the transformed up down method correctly, 'the expected proportion of positive responses is a monotonic function of stimulus level (at least over the region in which observations are made)' (pp 468, Levitt, 1971). If the psychometric function is not monotonic, then the reversals as determined by the adaptive procedure may not correctly asymptote, and an unreliable threshold may be reached. The method of constant stimuli, although taking more time, is not subject to such assumptions, and all parts of the data collected allow estimation of the psychometric function.

4.4.5 Construct validity: repeats and statistical confidence

A trial must be repeated a certain number of times in order to add statistical robustness to the result. If a question is asked just once, with a relatively small number of alternative forced choices, then chance plays a large role in determining the outcome. The greater the number of trials, the less chance will affect the result. Adaptive procedures terminate as a result of a predefined number of reversals, and so the exact number of trials per test is not the same, and use of the 1 up 1 down, or the 1 up 2 down methods may result in low numbers of trials per interval. Tests that utilised the method of constant stimuli also presented relatively low numbers of trials and in

addition the PMMA and the MBEA used trials that contained a mixture of interval changes, making it difficult to determine the number of repeats per interval.

The MCI does not state the minimum required number of repeats. In the study of Galvin *et al.* (2007) each participant repeated the MCI at least twice, with one participant repeating the MCI as many as 16 times. The average number of repeats used in this study was 6.2. The difficulty here is that each of the nine contours, although sharing the same interval, differ not only in their contour shape but also the interval from the first to the fifth note, e.g. 2 of the 9 contours (rising and falling) span four intervals, 6/9 contours (rising flat, falling flat, flat rising, flat falling, rising falling, falling rising) span two intervals and 1/9 of the contours (flat) span no intervals. When you break the 9 contours down in this way, there are fewer repeats than it at first appears.

The knowledge that the participant has about the test may also influence results: the MCI presents a visual response of the 9 contour options, and the participant can clearly see that the majority of trials consist of contours that change direction at the third note (e.g. falling flat, flat falling, rising flat, flat rising, rising falling and falling rising). For listeners that are struggling with this task, it is safer for them to choose one of those 6 as they are more likely to occur, because if some change of pitch is detected, the level of chance for the task goes from being one in 9, to one in 8 e.g. the listener knows that it cannot be the flat contour. The same is true if a rising pitch was detected, the listener is able to rule out all the falling flat combinations. Chance levels of 1/9 (11%) would only occur in this situation if the listener could not distinguish between the 9 contours at all.

4.4.6 Reliability

The reliability of the test is important to determine how well the test is measuring what it set out to measure, rather than measuring noise (Fitzpatrick *et al.*, 1998). If a test is not reliable, it cannot be deemed valid. There appears to be no predetermined criteria to state that the test can only be considered reliable if it has a reliability coefficient greater than a certain amount. An ICC of 0.82 can then be interpreted as meaning that 82% of the variability in measurement is due to genuine differences and 18% is due to methodology error (Bartlett and Frost, 2008). The level at which reliability is considered to be acceptable is a controversial issue; with many of the cut off points being arbitrary (Taber, 2017). Excellent reliability has been defined as > 0.75 (Fleiss, 1999), > 0.8 (Landis and Koch, 1977; Pinna *et al.*, 2007) and > 0.9 (Koo and Li, 2016). As such, for this work, excellent reliability was determined to be ≥ 0.8.

Intraclass correlation coefficients (ICC) can be used in place of the Pearson's product moment correlation coefficient. The ICC is beneficial over the traditional Pearson product moment

correlation coefficient because it enables an absolute agreement to be established rather than a close relationship. An example is given in McGraw and Wong (1996) using the pairs (0,4), (5,5) and (10,6). Conducting a Pearson's correlation on this data results in a coefficient of 1.00 whereas conducting the (the consistency) ICC results in a coefficient of 0.38. These differences can be put down to what is described by McGraw and Wong (1996) as linearity or additivity indices. The Pearson correlation uses a linearity index of agreement, as one variable can be related to another variable using a linear transformation ($y = ax + b$). The ICC consistency version uses an additivity index of agreement, relating one variable to another by adding a constant ($y = x + b$). The reason for this difference is due to underlying variances being different (the standard deviation of the first group is 2 and the second group is 5). Whilst the ICC is sensitive to differences in variance compared to the Pearson correlation, it is not appropriate to perform the ICC on groups with differing variances.

There are a number of different models of the ICC, and when deciding which ICC is appropriate, considerations involve: whether the model is 1 or 2 way – are there both row and column effects to consider; whether the column effects are considered random or fixed (described as Case 1, 2 or 3 by McGraw and Wong, 1996) – depending on the nature of the data in the columns; whether an interaction is present; whether the data is Type 1 (single measures e.g. body weight or single item scores) or Type *k* (average measures e.g. average of a number of scores); and whether the model should measure consistency (Type C, e.g. the models that do not include column variance or agreement (Type A, e.g. test retest reliability or agreement studies).

This is demonstrated in McGraw and Wong (1996, Table 6, p39) which shows 3 sets of paired data all with the same Pearson's *r* correlation of 0.67. McGraw and Wong state that when differences are small, *r*, ICC(C,1) and ICC(A,1) remain similar, however as the differences get bigger (and therefore agreement is reduced), this is reflected only in the ICC(A,1). It is important to use the absolute agreement (A) ICC rather than the consistency (C) of ICC when establishing test retest reliability, as larger mean differences between groups may not alter the Pearson's *r*, or the consistency ICC, however it is reflected in the agreement ICC. The ICC used throughout this thesis for test retest reliability was the two way random model, with single measures, and with absolute agreement, e.g. ICC (A,1).

4.4.7 Interpretation of results

In order to make a test interpretable for both the clinician and the test participant, it is important that the features and methodologies of the test are transparent. In addition it is important that any end results are presented in an appropriate context, for example if a test's final score is 75%, both clinician and test participant need to be aware of exactly what was tested, in order for a 75% score

to make sense in context. It is not immediately clear in the PMMA or in the MACarena what intervals were presented, and therefore final scores from these tests are not easy to interpret without doing further investigation into the stimuli, and that is not something that can be easily done in the timeframes of a clinical environment. Similar issues arise with the adaptive tests, especially as the 3 adaptive tests used in this thesis all employee different points of estimation along the psychometric curves (e.g. UW CAMP = 50%, SOECIC MTB PDT = 71%, MedEl MuSIC Test= 79%). Results from the MCI are also unclear: percentage scores from each of the intervals (1, 2, 3, 4, 5 semitones) might suggest that a certain score reflects ability at that interval, whereas because of the nature of the contours, the 1 semitone interval actually tests contours with pitch cues ranging from 2 to 4 semitones. The 5 semitone interval tests contours with pitch cues ranging from 10 to 20 semitones.

4.5 Gaps in knowledge

The literature surrounding pitch perception tests for CI users is lacking in strong evidence of validity and as such it is not known how suitable they are for use with CI users. Specifically, these gaps in knowledge have been summarised and will attempt to be addressed in this thesis:

1. What is the range of pitch perception ability in CI users?
2. What is the prevalence of non-monotonic psychometric functions for pitch perception in CI users?
3. Are the existing tests of pitch perception for CI users valid?
4. Are the existing tests of pitch perception for CI users reliable?
5. How can the measurement of pitch perception in CI users be improved?

Chapter 5 Experiment 1: Evaluating existing pitch tests

5.1 Introduction

A number of pitch tests have been designed and/or adapted for use with CI users. A range of methodologies and stimuli were used and it was not clear from the literature which test or tests might be considered most optimal for use with CI users. This chapter presents the initial experiment which reviews, analyses and evaluates pitch tests that have been designed for, or used with CI users. Each pitch test was initially evaluated using a sample of NHL, in order to establish how each test fares regarding a set of predefined evaluation criteria (section 5.1.2). The exception to this was the MCI, which was not included in the initial evaluation with NHL as at that time it was considered by this researcher to be more of a melody test than a pitch test. The results of this study were then used to determine which tests would be evaluated with a sample of CI users.

5.1.1 Aims

There were three aims to this study:

1. To establish whether these tests were suitable for use with CI users in terms of validity and reliability and clinical and functional use
2. To determine the best performing test (or tests) in the measurement of pitch perception in CI users
3. To establish which features should be utilised in future test design and which features should be abandoned, in order to create a valid, reliable and clinically functional test

5.1.2 Evaluation criteria and objectives

Based upon the ideal test qualities described in section 4.4, predetermined evaluation criteria were established in order to have a benchmark by which each test should be assessed. Definitions of the subheadings of validity were combined with the aspects from Fitzpatrick *et al.* (1998). Some overlap was seen, and as such, a combined list of evaluation criteria was determined.

Appropriateness of content	Do the test stimuli appropriately represent musical pitch? Is the difficulty level suitable for the typical range of CI users' abilities? Are floor and ceiling effects avoided?

Construct validity	Does the test provide a suitable number of repeats to ensure statistical confidence in the results? Is the methodology used to calculate the score suitable for use with CI users (e.g. does it take the possibility of non-monotonic psychometric functions for pitch interval perception into account)?
Criterion (concurrent) validity	Do test results correlate with theoretically similar tests, taken at the same time?
Reliability	Does the test produce a similarly repeatable score on retest?
Responsiveness	Does the test measure clinically relevant* changes?
Precision	Does the test measure a suitably small* interval?

** The appropriate magnitude of these may not be known*

As there is no 'gold standard' music perception test for CI users, it is difficult to assess concurrent validity by comparison to 'the best' existing, established and previously validated test. Instead, tests were compared to existing tests that were theoretically similar in their design.

The above was translated into the following research questions.

5.1.3 Research questions

1. Is the number of trial repeats enough to keep the likelihood of scores occurring by chance to less than 5%?
2. Are floor and/or ceiling effects present in the tests?
3. Do the tests show suitable reliability?
4. How do the test results compare with each other and is there any association between the test results?
5. Do the tests show significant differences between musicians and non-musicians?
6. Is the methodology used for calculating the result suitable for use with CI users?

5.1.4 Hypotheses

1 At least some of the tests will fail to provide enough repeats in order to achieve a high level of statistical confidence.
2 At least some of the tests will show floor and ceiling effects.
3 At least some of the tests will fail to show suitable levels of reliability on retest.
4 It is expected that large differences will be seen between tests
5 At least some of the tests will show significant differences between musicians and non-musicians.
6 At least some of the tests will have unsuitable methodology for CI users.

5.2 Methods

The method section describes the three parts to this Experiment I: Study 1 which compared pitch tests using NHL; Study 2 which assessed the test-retest reliability of the SOECIC MTB PDT using NHL; and Study 3 which compared pitch tests using CI users.

5.2.1 Materials - the pitch tests

Systematic searching of PubMed and Google Scholar was used initially, using the search terms 'music perception test cochlear implant users' and 'pitch perception test cochlear implant users' to obtain names of test batteries of music perception tests that have been used with CI users. Subsequent to that, each of the CI centres in the UK was contacted to see if any further music perception tests were known about or in use that had not been found using the initial search, however no further tests were found in this way. As pitch is a vital component of music, and in addition, because the majority of the music test batteries contained a measure of pitch perception, it was decided that this experiment would only compare pitch perception tests. The tests considered for evaluative review for this study are listed below, in chronological order:

1 The Primary Measures of Music Audiation (PMMA): tonal test (Gordon, 1979)
2 The Montreal Battery for the Evaluation of Amusia (MBEA): scale test (Peretz, Champod and Hyde, 2003)
3 The MACarena test platform amusia test: pitch test (Buechler, Lai and Dillier, 2004)

4 The Melodic Contour Identification test* (Galvin, Fu and Nogaki, 2007)

5 The University of Washington Clinical Assessment of Music Perception (UW CAMP): pitch test (Nimmons *et al.*, 2008)

6 The SOECIC Music Test Battery: pitch discrimination and identity test (van Besouw, 2010)

7 The MedEl Musical Sounds in Cochlear Implants (MuSIC) Test: pitch test (Brockmeier *et al.*, 2011)

** The MCI was not included within Study 1 (NHL) because at that time it had been overlooked as a pitch test, as it appeared initially to be more comparable to a melody test, however it was included for use in Study 3 (CI users).*

5.2.2 Study 1 - NHL pitch test comparison

This initial study aimed to compare the pitch perception tests using a group of NHL. Twenty three NHL were recruited to take part in all the above listed tests, with the exception of the MCI.

Equipment

Data collection was conducted in an acoustically treated room at the ISVR with ambient noise levels of <35 dB(A). All tests were run using a Dell Latitude E6400 laptop (with the exception of the MedEl MuSIC Test which was run using a Dell Latitude D610), running Windows XP, an external Behringer UCA202 soundcard and Behringer Truth B2031P loudspeaker, which was positioned 150cm from the position of the listener's head, with the tweeter at ear level. A flat screen monitor was positioned in front and slightly to one side of the participant so that they could see the graphical user interface for each test and respond using a computer mouse.

Calibration

Stimuli for all experiments were presented in close to 'free field' conditions. Due to the 'behind the ear' processor of the CI, Bamford Kowal Bench (BKB) sentence testing has typically and traditionally been tested in this way, avoiding the use of headphones. Although it is possible to provide supra-aural headphones that surround the BTE processor of the CI, this may not always be possible or comfortable for CI users. In order to create a similar and familiar clinical testing environment, a similar set up to BKB sentence testing was used for Experiment 1.

The aim was to present all stimuli at the same loudness level, 65 dB(A), which is the level used with BKB sentence testing. Only three of the pitch tests provided their own calibration tone (UW CAMP,

SOECIC MTB PDT, and MedEl MuSIC Test). Prior to the first test session, each test was completed by the experimenter in order to subjectively determine that the stimuli throughout the test were close to 65 dB(A). A hand held Bruel and Kjaer Type 2235 Precision sound level meter (SLM) was mounted on a tripod at the position of the listener's ear (in their absence). Due to the time varying nature of musical stimuli, these levels were variable, however the experimenter was satisfied with levels that fell between 60-70dB(A) for the duration of the test by using listening checks and the hand held SLM.

Participants

The 23 NHL (13 female, 10 male, 11 musicians[1], 12 non-musicians, mean age 28 years, ±6 SD) were recruited via opportunity sampling from the University of Southampton, were aged between 19 and 43, and all had hearing thresholds in quiet of 20 dB HL or better, and reported no amusia or tone deafness, prior to taking part in the study. Pure tone audiometry was conducted if it had not been done 6 weeks prior to testing. Testing took place between October 2010 and February 2011. Participants were not paid, however they were offered confectionary plus information about their pitch perception performance on each test.

Ethical approval

Ethical approval was applied for and approved by the ISVR Human Experimentation Safety and Ethics Committee (reference 1135) and the University of Southampton's (UOS) Research Governance Office (reference 7511).

Procedure

Normal hearing listeners attended for one or two sessions (not all participants were able to stay long enough to complete all tests in the first session) and took part in 6 tests of pitch perception. Order of testing was randomised across participants, using a Latin Square design, in order to minimise order effects and learning effects, in a similar way to the example below.

[1]*Musician* defined by self-report: if any participant confirmed that they held any musical qualification, or considered themselves to be a musician, they were considered to be a musician for this thesis.

NHL 1	A	B	C
NHL 2	B	C	A
NHL 3	C	A	B

On completion, the results were explained to each participant, and any comments were noted.

5.2.3 Study 2 - NHL SOECIC MTB PDT reliability analysis

This second study aimed to assess the test retest reliability of the SOECIC MTB PDT, the only test that did not show ceiling effects with NHL. Eighteen NHL completed the SOECIC MTB PDT at T1 and T2, in order to analyse the reliability. Ceiling effects plus time limitations meant that testing reliability in Study 1 would not have been feasible. The SOECIC MTB PDT was the only test of the 6 that did not show any ceiling effects. This study was conducted in the UK and in Belgium.

Equipment

Data collection in the UK was conducted within a sound treated room within the ISVR, UOS with ambient noise levels of < 35 dB(A), which was tested prior to conducting the pitch tests, using a hand held Bruel and Kjaer Type 2235 Precision SLM. In Belgium, data was collected in a quiet office, within the Cochlear Technology Centre (CTC). The SOECIC MTB PDT was run using a Dell Latitude E6400 laptop running Windows XP. An external soundcard (Behringer UCA202) was used to connect the laptop to supra-aural headphones (Senheisser HD 280 Professional).

Calibration

Headphones were used in this study for ease of calibration whilst testing across two sites (CTC and ISVR), and because this study did not use CI users, and was only interested in comparing results from T1 and T2, the issues regarding headphone use (described above) did not impact this study. The headphones and laptop were calibrated using an artificial ear with a flat plate coupler, to ensure that the output of the tones presented with the SOECIC MTB PDT ranged between 60-70 dB(A). Participants responded to the graphical user interface using a computer mouse.

Participants

The 18 NHL (6 female, 12 male, 9 musicians[2], 9 non-musicians) were recruited via opportunity sampling from the University of Southampton, UK, and CTC, Mechelen, Belgium. These NHL consisted of 11 self-reported normally-hearing adults, recruited from Belgium (time and facilities were not available to perform pure tone audiometry at time of testing) and seven NHL from the UK (four of whom had already taken part in Study 1). Pure tone audiometry was conducted if it had not been done 6 weeks prior to testing (for NHL in the UK). Contraindications were self-reported amusia. Testing took place between November 2011 and February 2012.

Ethical approval

Ethical approval was applied for and approved by the ISVR Safety and Ethics Committee (reference 1135).

Procedure

All participants attended for one session where they completed the SOECIC MTB PDT twice. On completion, the results were explained to them.

5.2.4 Study 3 - CI pitch test comparison

Equipment and calibration as in Study 1.

Participants

The 15 CI users (7 female, 8 male, 2 musicians[3], 13 non-musicians, mean age 67 years, ±16 SD) were recruited via written invitation using the database at University of Southampton Auditory Implant Service (USAIS, formerly SOECIC). The general inclusion criteria for the study were that potential recruits had to be adults with one or two CI and be resident in mainland UK. USAIS is associated with a large number of research projects and as such has methods in place to ensure that patients are not over invited to participate in research. This meant that certain extra exclusion

[2]*Musician* defined by self-report: if any participant confirmed that they held any musical qualification, or considered themselves to be a musician, they were considered to be a musician for this thesis.

[3]*Musician* defined by self-report: if any participant confirmed that they held any musical qualification, or considered themselves to be a musician, they were considered to be a musician for this thesis.

criteria were included, for recruitment purposes only: patients that had a switch on date of less than 12 months prior to recruitment were not included. In addition, patients that were considered by the research coordinator to be unlikely to cope well with being invited to, or participating in the research, were not invited.

Table 5.1 provides demographic details of the CI users. BKB scores ranged from 34 – 99%. Length of deafness ranged from 2 years to 73 years (mean 36 years, SD 25 years). Two of the 15 CI users (CI 6 and CI 11) were bilaterally stimulated: CI 6 was implanted with the Neurelec device which provides bilateral stimulation and CI 11 was visually impaired, and therefore had been implanted with two CIs. Four CI users also had a contralateral hearing aid (HA), and in order to occlude any residual hearing, the HA was switched off, but the HA and ear mould were left in the contralateral ear. Testing took place between December 2011 and April 2012.

Table 5.1 CI user demographics for Experiment 1

ID	Age (years)	Sex	Pre/post lingual deafness	CI manufacturer	Listening mode	Duration post initial tuning (months)	Prior music training
CI1	59	M	Post	Cochlear	Unilateral	96	Yes
CI2	64	M	Post	Cochlear	Unilateral	204	No
CI3	88	F	Post	Cochlear	Unilateral	216	Yes
CI4	79	M	Post	MedEl	Unilateral	24	No
CI5	83	M	Post	Cochlear	Unilateral	96	No
CI6	74	M	Post	Neurelec	Bilateral	12	No
CI7	91	M	Post	Advanced Bionics	Unilateral	12	No
CI8	51	F	Pre	Cochlear	Unilateral	12	No
CI9	66	M	Post	Advanced Bionics	Unilateral	24	No
CI10	38	M	Post	Advanced Bionics	Unilateral	12	No
CI11	56	F	Post	Advanced Bionics	Bilateral	24	No
CI12	52	F	Post	Cochlear	Unilateral	12	No
CI13	69	F	Post	MedEl	Unilateral	12	No
CI14	76	F	Post	MedEl	Unilateral	288	No
CI15	58	F	Pre	MedEl	Unilateral	12	No

Ethical approval

Ethical approval was obtained from the National Research Ethics Service (NRES) reference number 11/SC/0263 and from the UOS Institute of Sound and Vibration Research Safety and Ethics Committee, references 1163 and 1214 and the UOS's Research Governance Office, reference 7940. Informed written consent was taken from all participants prior to proceeding.

Procedure

Participants attended for one 3 hour session and took part in 5 tests of pitch perception, a total of two times each. Participants were asked to sign the consent form and the study was explained to them. Participants were asked to complete a few questions about their musicianship status. Participants were asked to use their 'everyday' program on their CI, and if they wore a contralateral HA, were asked to continue to wear it, but switch it off, during testing. A Latin Square was used to determine the order of tests (PMMA, MedEl MuSIC Test, SOECIC MTB PDT and UW CAMP and MCI) for each participant, in order to minimise any order and learning effects. The MACarena and the MBEA were not used for Study 3 due to their expected poor performance with CI users. Prior to each pitch test, instructions were given and the researcher clarified with each participant that they understood the task. Throughout testing, a document 'Instructions for participants' was available to them (Appendix A). At the end of each test, participants had their results explained to them.

5.3 Initial Results

Initially, each test was set up and used repeatedly by the experimenter, in order to look at the stimuli content, the number of trials, what the test was able to measure and how the results were presented. These results are presented in Table 5.2 and Table 5.3.

The adaptive method tests (MedEl MuSIC Test, UW CAMP, and SOECIC MTB PDT) were then completed by the experimenter, with the experimenter aiming to get perfect or near perfect performance (e.g. every trial correct), in order to establish the algorithm rules, to see how the adaptive procedures responded and to investigate the calculation of the final result.

Finally, each test was investigated regarding the number of trials that it presented, and how many trials might be necessary in order to obtain a statistically robust result, given the level of chance for each test, and the rules that governed how each test proceeded.

5.3.1 Initial test comparison

Table 5.2 Comparison of test details for the MCS tests

Name and date	Stimuli & Range	Interval size	Test type, number of trials, chance level
Primary Measure of Music Assessment (PMMA) v1.0 Tonal test Gordon (1979)	'electronically produced' 1 harmonic tone 262 Hz (C4) – 523 Hz (C5)	Tests intervals from 1 – 12 semitones, but each level is not distinguishable as more than one interval tested in each trial	Same/different 40 x 2-5 tone phrase discrimination 20/40 trials different Chance = 50%
MACarena Amusia test of pitch Buechler, Lai and Dillier (2004)	MIDI piano 262 Hz (C4) – 293.66 (D4)	Tests intervals from 1 – 2 semitones	24 x 2AFC tone discrimination 16/24 trials different Chance = 50%
Montreal Battery for the Evaluation of Amusia (MBEA) Scale test Peretz, Champod and Hyde (2003)	MIDI piano 165 Hz (E3) – 699 Hz (F5)	Tests intervals of 1 – 6 semitones	31 x 2AFC phrase discrimination 15/31 trials different 1/31 trial oddball Chance = 50%
Melodic Contour Identification (MCI) Galvin, Fu and Nogaki (2007)	Synthesized complex tone, 3 harmonics 440 Hz (A4) – 1397 Hz (F6)	1 semitone interval tests 2 – 4 semitones, 5 semitones tests 10 – 20 semitones	5 x 27 9AFC contour discrimination Chance = 11%

Table 5.3 Comparison of test details for the adaptive tests

Name and date	Stimuli and range	Intervals tested and step size	Test details and final score
MedEl MuSIC Test: Pitch Test Brockmeier *et al.* (2011)	Real recorded piano tones 261.63 Hz (C4) – 4186 Hz (C8)	Can test intervals from 0.5 – 48 semitones (when target note set at C4) Starts at 11 semitones above C4 (261.63 Hz) Initial step: 5.5 semitones Decreases to 3, 1.5, 1 and 0.5 semitones	2AFC adaptive 3 down 1 up, tracking at 79% Terminates at 8 reversals, final score is approximated by average of final 5*
UW CAMP Pitch test Nimmons *et al.* (2008)	Synthetic complex tone with spectral envelope from piano at C4, uniform temporal envelope, summed sine waves 261.63 Hz (C4) – 784 Hz (G5)	Can test intervals from 1 – 12 semitones Starts at 12 semitones above 'target note' (C4, E4, G4) Initial step: 6 semitones Subsequent step sizes: 3, 2, 1 semitones	2AFC adaptive 1 down 1 up, tracking at 50% Terminates at 8 reversals, final score is average of final 6 Averages 3 of 3 base notes (C4,E4,G4) interweaved staircase
SOECIC MTB PDT Pitch discrimination test van Besouw (2010)	Synthesized sine tone Possible range: 87.3 Hz (F2) - 3520 Hz (A7) 'all frequencies' for Experiment 1: Study 1 'F4' for Experiment 1: Study 2 and 3 and Experiment 2	Can test intervals from 0.01 – 16 semitones Starts at 16, 8 or 4 semitones (randomly), can be either higher or lower than chosen target note. Step sizes are predetermined depending on where the algorithm is.	3AFC procedure 2 down 1 up tracking at 71% Terminates at 7 reversals, final score is average of final 5

** this estimates the final score well, however this is not the true algorithm and although numerous attempts have been made to contact the originators of the test, no explanation has been forthcoming. Several attempts at decoding the author's own data and data of the participants has also not lead to anything more accurate than the estimation of 5 of 8 reversals, even allowing for large rounding errors.*

5.3.2 Algorithm rules, termination and final score calculation

This section will explain the rules for the calculation of the results of each of the adaptive pitch perception tests.

MedEl MuSIC Test

The MedEl MuSIC test uses a 3 up 1 down adaptive staircase procedure, meaning that participants must achieve 3 correct scores in a row before the interval size is made more difficult by becoming smaller. It starts at 22 quartertones (11 semitones) above the base note (C4, 261.63 Hz). These are the default settings, however they can be adjusted.

For perfect performers, the intervals descend by 11 quartertones (5.5 semitones) and then 10 quartertones (5 semitones), therefore presenting C4 + 22 quartertones three times, then C4 + 11 quartertones three times and finally C4 + 1 quartertone three times. After this performance, with all responses correct, the test terminates and the final score is 1 quartertone (0.5 semitone). This can be seen in Figure 5.1. There is no ascension of the staircase within a perfect performance.

For non-perfect performers (Figure 5.2), the initial interval is 11 quartertones (5.5 semitones), from the starting note, until the third reversal. If the participant cannot score correctly 3 times in a row, then this interval is maintained throughout the test. At the third reversal, the interval size for ascension and descension of the staircase drops to 6 quartertones (3 semitones) and is maintained until the 5th reversal. The interval size then drops to 3 quartertones (1.5 semitones) until the 7th reversal and finally drops to 2 quartertones (1 semitone) for the final ascension or descension of the staircase and then the staircase is terminated at the 8th reversal.

The staircase terminates under 3 circumstances. The first is in the case of perfect performance, where the staircase terminates once three correct responses at 1 quartertone have been recorded. The staircase will also terminate if the maximum interval (96 quartertones, or 48 semitones, when the starting note is C4) is incorrectly answered. Finally, and most commonly, termination of the staircase occurs after 8 reversals.

The final score in the first example of perfect performance is 1 quartertone (0.5 semitone). The final score in the second example, where the participant has answered incorrectly at the largest interval of 96 quartertones, is scored as 0 quartertones, to indicate that the test has 'failed' to score.

There are no published documents that explain the calculation of the final score when the MedEl MuSIC test terminates after 8 reversals. The author of this thesis attempted to determine the rules

behind the calculation of the final score, using the data from both NHL and CI users from later in this chapter. Combinations of the final 2, 3, 4, 5, 6, 7 and 8 reversals were analysed. Means, medians and modes were all calculated, however no reliable rule could be found that explained the final score calculated by the test itself.

The original author of the paper (Hanna Brockmeier) and the originator of the MedEl MuSIC test package (Denis Fitzgerald) were contacted and neither were able to provide any further details. The final score produced by the MedEl MuSIC test can be approximated by averaging the final 5 of the 8 reversals, however even allowing for large rounding errors, this can only give an approximate value.

UW CAMP

The UW CAMP uses a 1 up 1 down adaptive staircase procedure, meaning that the staircase will ascend as soon as a participant gets one response incorrect and will descend as soon as they get one response correct. It starts at 12 semitones above each of the three starting notes: C4 (261.63 Hz), E4 (329.63 Hz) and G4 (392 Hz). Each starting note run is repeated 3 times.

For perfect performers, the intervals descend by 6 semitones, then 3 semitones, then 2 semitones, then remain at 1 semitone until 8 reversals are complete. This can be seen in Figure 5.3. Although the smallest interval presented is 1 semitone, the UW CAMP adds an 'automatic reversal' at '0 semitones' in order to satisfy the necessity of the 8 reversals. As such, the final score is reported by the test software as 0.5 semitone (the average of the final 6 reversals: 0, 1, 0, 1, 0, 1).

For non-perfect performers, the intervals also follow the pattern of reducing from 6, to 3, to 2 and then finally 1 semitone throughout the descension and ascension. The UW CAMP cannot increase the interval sizes above the +12 semitones above each of the starting notes, therefore if the participant cannot respond correctly to the first descent comparing +12 with +6 semitone interval, then the interval sizes then continue at 1 semitone throughout the test. This can be seen in the third run within Figure 5.4. The staircase always terminates after 8 reversals. The final score is calculated by taking the average of the final 6 reversals.

SOECIC MTB PDT

The SOECIC MTB PDT uses a 2 up 1 down adaptive staircase procedure, meaning that participants must achieve 2 correct scores in a row before the interval size is made more difficult by becoming

smaller. The target note can be chosen within the settings of the test, and the test then randomly decides whether the initial comparison will be 16, 8 or 4 semitones above or below the target note.

Perfect performance was not achievable because the SOECIC MTB PDT tests down to 1 cent which is below the capabilities of most NHL. The intervals always descend or ascend using the predetermined intervals of 16, 8, 4, 2 and 1 semitones, and then the intervals drop to 0.64, 0.32, 0.16, 0.08, 0.04, 0.02 and 0.01 semitone. If a participant does not get the initial interval correct at 4 or 8 semitones then the interval will increase, however no increases are possible above 16 semitones. The staircase always terminates after 7 reversals, and the final score is calculated by taking the average of the final 5 reversals.

5.3.3 Perfect or near perfect performance

The MedEl MuSIC Test requires that listeners obtain 3 correct responses in a row (3 down 1 up) before the interval size is reduced. This test terminates after 8 reversals, however with perfect performance, as shown below, the test also terminates. The final score is then = 0.5 semitones (reported as a score of 1 quartertone).

11 11 11

5.5 5.5 5.5

0.5 0.5 0.5 (test terminates)

The UW CAMP test requires that listeners only need to get one correct response (1 down 1 up) before the interval is reduced. This test terminates after 8 reversals (indicated by *), and in order to do this with perfect performance, an automatic reversal at '0 semitones' is added. The final score is calculated by taking the mean of the final 6 of 8 reversals, and the final score is calculated as 0.5 semitones.

12

6

3

1 1* 1* 1* 1*

0* 0* 0* 0* (test terminates)

Another example of good performance with the UW CAMP, as shown below, leads to a final score of less than 1 semitone (final score = 0.83 semitones).

12

6

3 2* 2*

1* 1* 1 1* 1*

0* 0* (test terminates)

Both of the UW CAMP examples above are misleading, as the smallest interval that the test measures is 1 semitone. Any scores that equal 1 or below should therefore be interpreted as a final score of 1 semitone, although this isn't made clear in either the 2008 paper or when the score is delivered to the test user at the end of the test. Nimmons *et al.* (2008) state 'For some listeners, the raw pitch score and the psychometric curve suggest that the true DL was less than 1 semitone, but because 1 semitone was the smallest tested interval, it was considered the best achievable score' (p5).

The SOECIC MTB PDT requires that listeners need to get two correct responses (2 down 1 up) in a row before the interval is reduced. Excellent performance is shown below, and the test terminates once 7 reversals (indicated by *) have occurred. The final score is calculated by taking the mean of the final 5 of 7 reversals, and the score is 0.06 semitones.

16 16

8 8

4 4

2 2

1 1

0.64 0.64

0.32 0.32

0.16 0.16

0.08 0.08 0.08 0.08* 0.08 0.08* 0.08 0.08*

0.04* 0.04* 0.04* 0.04*

(test terminates)

5.3.4 The role of chance

The role that chance plays in each test was investigated by calculating probabilities based upon number of trials and the level of chance. This was much more straightforward to do for MCS tests because of their fixed number of trials. For the adaptive tests, each test terminates depending on the participant's performance therefore the number of trials cannot be predetermined, however chance levels associated with each adaptive staircase algorithm were investigated here. Specific examples using individual NHL and CI results are presented in sections 5.4.1 and 0.

The binomial distribution was used as a model for determining the chance levels surrounding success or failure in the pitch tests because an answer could either be correct or incorrect. An acceptable level of chance was taken as $p < 0.05$, which is the traditional alpha value used in statistics.

Levels of chance associated with numbers of trials and numbers of successes can be calculated by multiplying each probability by itself for each repeat, e.g. to work out the probability of scoring 3 successes in a row, with a probability of 0.5:

0.5 x 0.5 x 0.5 = 0.125

This is therefore the probability for both:

3/3 success and 0/3 success

The probability of scoring either 2/3 or 1/3 should be equal, and therefore 1- (2 x 0.125) = 0.75

So probability of each of 1/3 and 2/3 is 0.875 / 2 = 0.375

Successes/trials	PROBABILITY	cumulative probability (that score or higher)	
0/3	0.125	1	e.g. 0, 1, 2 or 3/3
1/3	0.375	0.875	e.g. 1, 2, or 3/3
2/3	0.375	0.5	e.g. 2 or 3/3
3/3	0.125	0.125	e.g. only 3/3

These probabilities and cumulative probabilities were calculated using an online binomial calculator (http://stattrek.com/online-calculator/binomial.aspx)

This approach was used to calculate the minimum number of successful trials for each of the MCS tests in order for the likelihood of success purely due to chance to be considered to be lower than the alpha value of $p = 0.05$.

The PMMA had chance levels of 50% due to the question being 'same/different' e.g. 2 AFC. Therefore with 40 trials, a minimum of 26 correct trials was needed in order to achieve $p < 0.05$, e.g. $p = 0.04$.

The MACarena had chance levels of 50% due to the question being 'same different' e.g. 2 AFC. Therefore with 24 trials, a minimum of 17 correct trials was needed in order to achieve $p < 0.05$, e.g. $p = 0.03$.

The MBEA had chance levels of 50% due to the question being 'same different' e.g. 2 AFC. Therefore with 31 trials, a minimum of 21 correct trials was needed in order to achieve $p < 0.05$, e.g. $p = 0.04$.

The MCI had chance levels of 11%, e.g. 9 AFC. Therefore, with 27 trials, a minimum number of 7 correct trials was needed in order to achieve $p < 0.05$, e.g. $p = 0.03$.

This approach was also used to calculate the levels of chance associated with 3 down, 1 up, 2 down 1 up and 1 down, 1 up staircase methods, as used in the MedEl MuSIC Test, SOECIC MTB PDT and UW CAMP tests.

The MedEl MuSIC Test was 2 AFC and chance = 50% and required 3 correct responses in a row in order to descend the staircase. The probability of this occurring by chance is 0.125, which breaches the cut off of 0.05. This means that a 2 AFC 3 down 1 up adaptive procedure has the likelihood of success due to chance alone to a greater extent than is allowed in statistics generally. As shown below, it would require 5 correct trials out of 5 in order to reduce the chance level (of each descent of the staircase) to below 0.05.

No of successes(x)	no of trials	probability of X=x
3	3	0.125
4	4	0.0625
5	5	0.03125 (< 0.05)

The UW CAMP test was 2 AFC with chance = 50%, and required 1 correct response in order to descend the staircase. The probability of this occurring by chance is 50%, which is much higher than the cut off of 0.05. As in the example above, it also requires 5 correct trials out of 5 in order to reduce the chance level (of each descent of the staircase) to below 0.05.

The SOECIC MTB PDT was 2 AFC with chance = 33%, and required 2 correct responses in a row in order to descend the staircase. The probability of this occurring by chance is 0.11, which is still higher than the cut off of 0.05. Increasing the adaptive procedure to a 3 down 1 up staircase would reduce chance levels (of each descent of the staircase) to 0.04.

No of successes(x)	no of trials	probability of X=x
3	3	0.04 (< 0.05)

This section has described the basic features of the tests used in this experiment and has presented how the adaptive tests respond to a hypothetical perfect performance, and it has also demonstrated how chance might affect progression of the adaptive tests, and their final scores. The next section will present the results with NHL.

5.4 NHL Results

This section presents the results of the MACarena, the MBEA, the PMMA, the MedEl MuSIC Test, the UW CAMP and the SOECIC MTB PDT, using a group of 23 NHL. Results from T1 only were analysed; this was to preserve the face validity of the investigation, as each clinical test is typically only used once (e.g. T1), rather than twice, given the time constraints of a busy clinic. Data from T2 was collected to investigate reliability only. Data from the UW CAMP was adjusted so that any scores of < 1 were reassigned the value of 1, except for when testing reliability on retest.

Each test was analysed to determine whether the number of trials was sufficient to keep the by chance alone to less than 5%; whether floor and ceiling effects affected the data; comparisons were made between the tests; and sensitivity to musicianship was investigated.

Data distribution was found to be significantly non-normal at the $p < 0.05$ level for all tests except for the MBEA, and as such, nonparametric statistical approaches were used.

Table 5.4 NHL median, interquartile range (IQR), maximum (max) and minimum (min) scores for all tests

Scores in semitones or %, n = 23

Test	Median	IQR	min	max
PMMA (%)	100	7.5	77.5	100*
MACarena (%)	100	4.65	71.9	100*
MBEA (%)	90.32	11.29	70.97	100*
MedEl MuSIC Test (semitones)	0.5	0.5	0.5*	11
UW CAMP 262 Hz (semitones)	1	0.03	1*	4.89
UW CAMP 330 Hz (semitones)	1	0	1*	3.22
UW CAMP 392 Hz (semitones)	1	0.09	1*	4.28
SOECIC MTB PDT (semitones)	0.29	0.25	0.11	2.8

** at ceiling*

5.4.1 Trial number and the role of chance in adaptive methods

Due to the nature of the descending staircase seen in the adaptive methods, some individual intervals may not have a very large number of trials. Using the chance levels for each test, the binomial calculator was used to determine whether actual examples of typical progression of the adaptive procedure offered enough trials in order to minimise the effect of chance. These calculations were not undertaken for every participant, however two examples were taken for each adaptive test, a good performer and a poor performer.

MedEl MuSIC Test: 2 AFC, chance = 50%, 3 down 1 up, no repeats

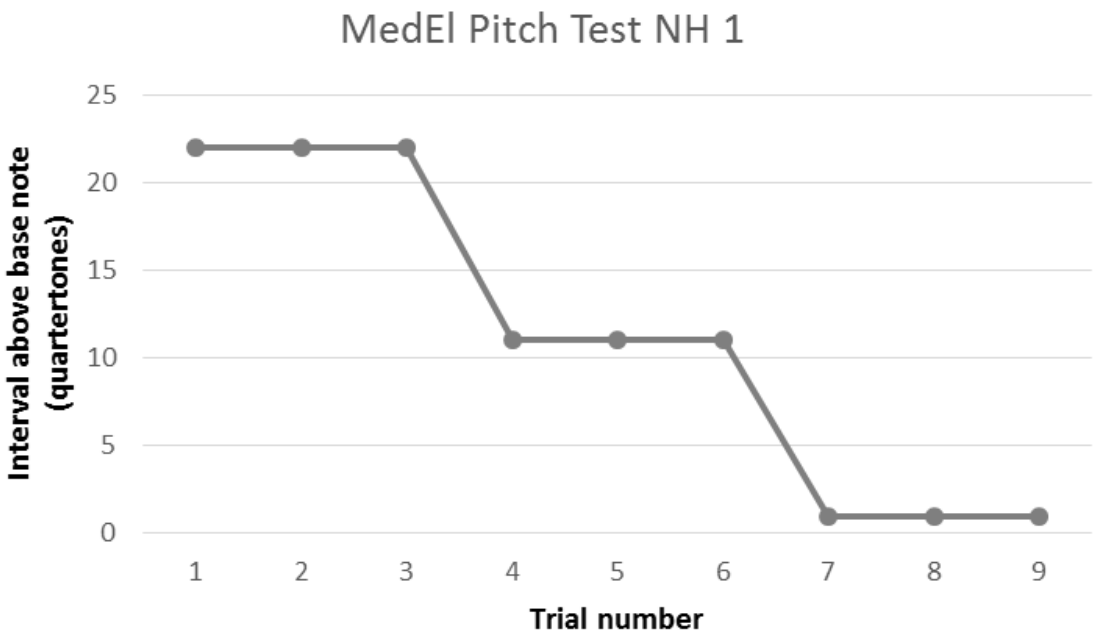


Figure 5.1 Example of a good performer (NHL 1) using the MedEl MuSIC Test, showing the adaptive staircase. Final score was 1 quartertone (0.5 semitone), however there were insufficient trials within this test (and with this level of performance) to ensure that the final score had $p < 0.05$.

In the example above, the final score was 1 quartertone (0.5 semitone) as the participant was successful on every interval.

Below is a breakdown of NHL 1's performance, for each interval presented, the number of successes and the number of trials. The MedEl MuSIC Test used intervals measured in quartertones, which is half a semitone. It also presented the final score in quartertones, and so that is how it is presented here, for clarity.

22 quartertones: 3/3 successful

11 quartertones: 3/3 successful

1 quartertone: 3/3 successful

As shown in section 5.3.4, when the level of chance is 50%, 5 trials (or more) are needed to keep $p < 0.05$, and as such, there were not enough successful trials presented for 22, 11 and 1 quartertone in this example. The final score was 1 quartertone, and due to the low number of trials, the likelihood that this was due to chance is high.

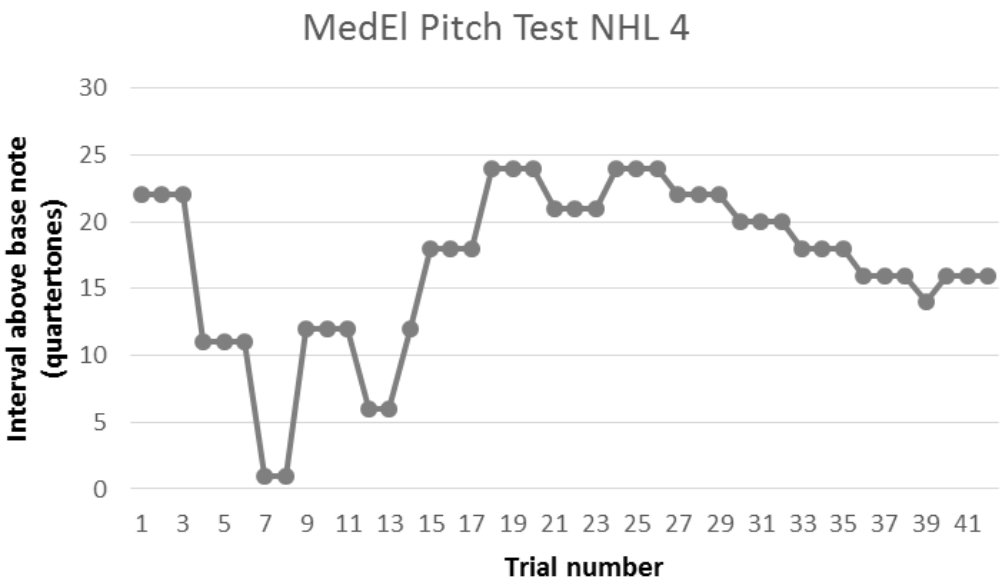


Figure 5.2 Example of a poor performer (NHL 4) using the MedEl MuSIC Test, showing the adaptive staircase. Final score was 18 quartertones (9 semitones), and this test (with this level of performance) had sufficient trials to ensure that the final score had $p < 0.05$.

In the example above, the final score was 18 quartertones (9 semitones). The final score of the MedEl MuSIC Test is approximated by averaging the final 5 of the 8 reversals. Figure 5.2 shows the final 5 reversals (in reverse order) as 16, 14, 24, 21 and 24, and their average is 19.8 quartertones, a close approximation to the final score obtained by the test, of 18 quartertones.

Below is a breakdown of NHL 4's performance. Asterisks represent a suitably high number of successes for $p < 0.05$:

24 quartertones: 6/6 successful*

22 quartertones: 6/6 successful*

21 quartertones: 2/3 successful

20 quartertones: 3/3 successful

18 quartertones: 5/6 successful

16 quartertones: 6/6 successful*

14 quartertones: 0/1 successful

12 quartertones: 3/4 successful

11 quartertones: 3/3 successful

6 quartertones: ½ successful

1 quartertone: ½ successful

This indicates that enough successful trials were presented for 24, 22 and 16 quartertones. The final score was 18 quartertones, which didn't have enough trials, however it is close to the interval 16 quartertones, and so the likelihood that this was due to chance is small.

UW CAMP Pitch Test: 2 AFC, chance = 50%, 1 down 1 up, 3 repeats

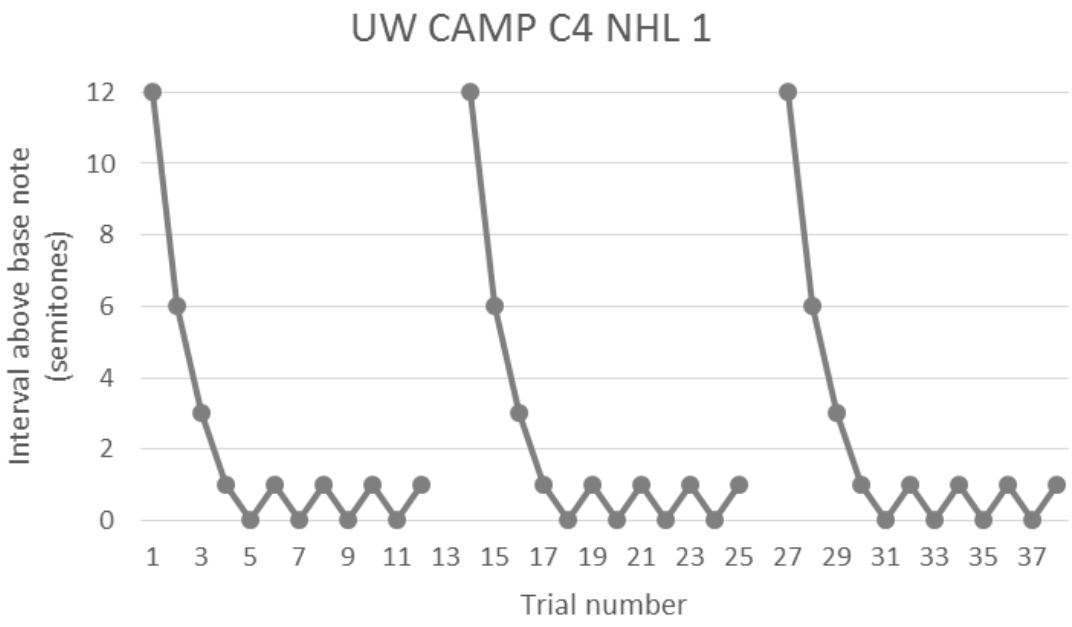


Figure 5.3 Example of a good performer (NHL 1) using the UW CAMP Test, showing the adaptive staircase. Final score was '0.5' semitones (1 semitone) (run 1: 0.5, run 2: 0.5, run 3: 0.5) and this test (with this level of performance and three repeats) had sufficient trials to ensure that the final score had $p < 0.05$.

In the example above, the final score was 0.5 (1) semitone. The final score of the UW CAMP is calculated by averaging scores from each of the 3 runs, and these are calculated by averaging the final 6 of the 8 reversals for each run. Figure 5.3 shows identical staircases for all 3 runs, and the final 6 reversals (in reverse order) for all runs were 1, 0, 1, 0, 1, 0, and their average is 0.5 semitone.

Below is a breakdown of NHL 1's performance. Asterisks represent a suitably high number of successes for $p < 0.05$:

12 semitones: 3/3 successful

6 semitones: 3/3 successful

3 semitones: 3/3 successful

1 semitone: 15/15 successful*

This indicates that the only interval with enough successful trials was 1 semitone. However, as the final score was 1 semitone, the likelihood that this was due to chance is small.

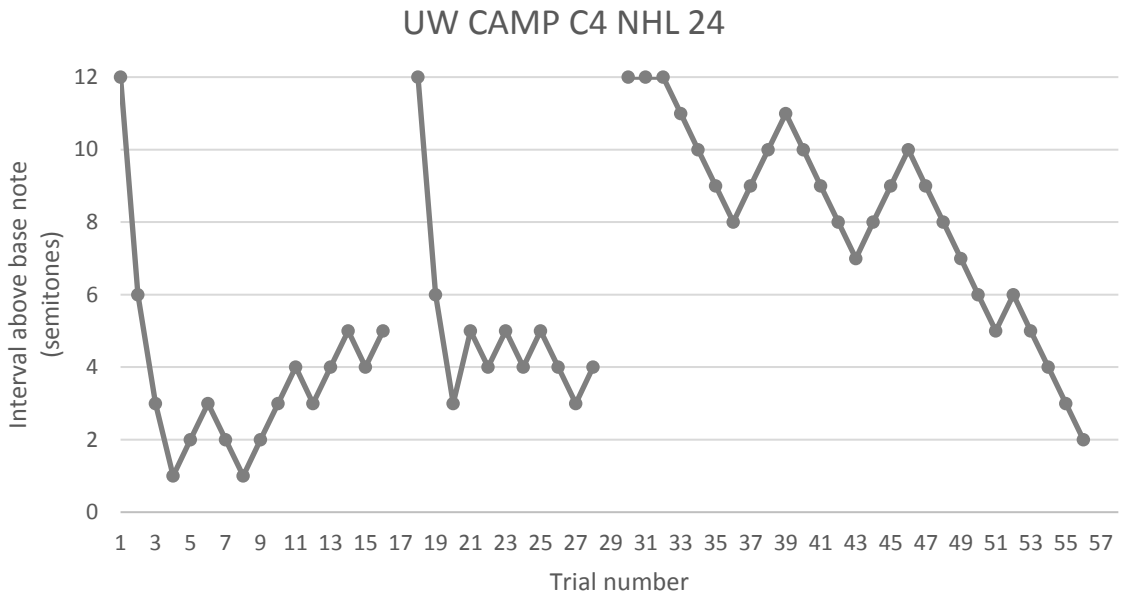


Figure 5.4 Example of a poor performer (NHL 24) using the UW CAMP Test, showing the adaptive staircase. Final score was 4.89 semitones (run 1: 3.67 run 2: 4.17 run 3: 6.83) and this test (with this level of performance and three repeats) had sufficient trials to ensure that the final score had $p < 0.05$.

In the example above, the final score was 4.89 semitones. Figure 5.4 run 1 shows the final 6 reversals (in reverse order) as 5, 4, 5, 3, 4 and 1, and their average is 3.66. Run 2 shows the final 6 reversals as 4, 3, 5, 4, 5 and 4 and their average is 4.17. Run 3 shows the final 6 reversals as 2, 6, 5, 10, 7 and 11 and their average is 6.83. The average of these 3 run final scores is 4.89 semitones.

Below is a breakdown of NHL 24's performance. Asterisks represent a suitably high number of successes for $p < 0.05$:

12 semitones: 3/5 successful

11 semitones: 2/2 successful

10 semitones: 3/4 successful

9 semitones: 3/5 successful

8 semitones: 2/4 successful

7 semitones: 1/2 successful

6 semitones: 3/4 successful

5 semitones: 5/7 successful*

4 semitones: 3/8 successful

3 semitones: 3/7 successful

2 semitones: 1/4 successful

1 semitones: 0/2 successful

This indicates that enough successful trials were presented for only 5 semitones. The final score was 4.89 semitones, which is close to the interval 5 semitones, and so the likelihood that this was due to chance is small.

SOECIC MTB PDT: 3 AFC, chance = 33%, 2 down 1 up, no repeats

Data could not be analysed for the SOECIC MTB PDT results with NHL as the version of the SOECIC MTB PDT used for this group did not automatically store the adaptive staircase procedure.

5.4.2 Floor and ceiling effects with NHL

For the 3 MCS tests (e.g. MACarena, MBEA and PMMA), ceiling effects were defined as scores of 100%, and scores at chance level (50%) were considered to be floor effects.

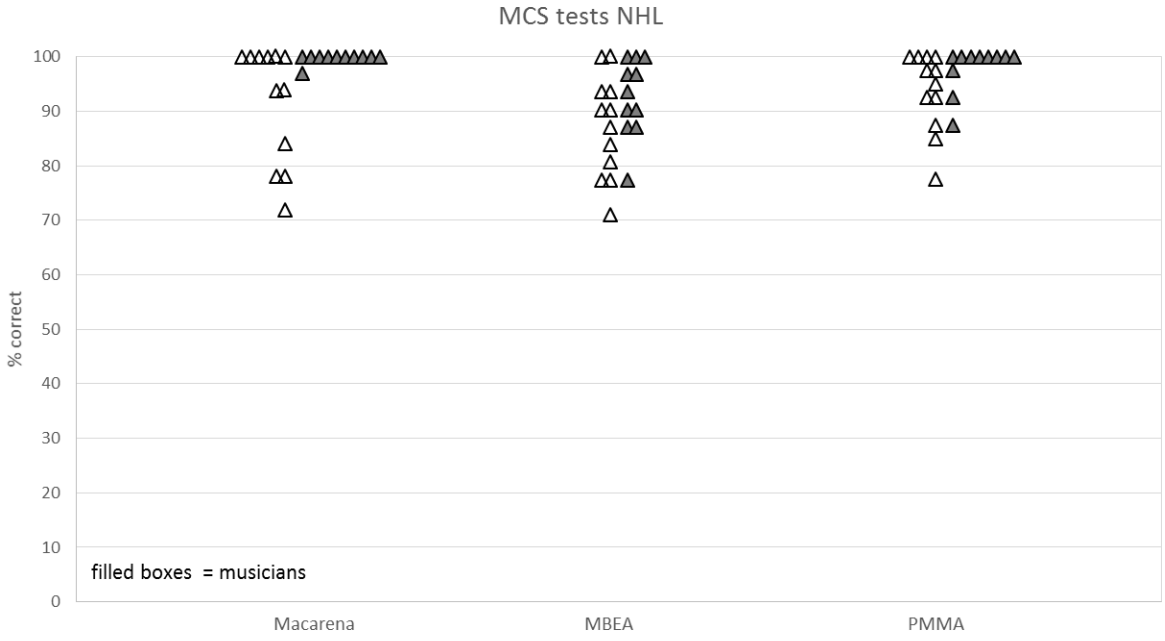


Figure 5.5: Pitch tests using the method of constant stimuli with normal hearing listeners (time 1 data only) showing ceiling effects. White triangles = non musicians, grey triangles = musicians.

MACarena: ceiling effects (100%) were seen for 16/23 NHL. No floor effects (~50%) were seen.

MBEA: ceiling effects (100%) were seen for 5/23 NHL. No floor effects (~50%) were seen.

PMMA: ceiling effects (100%) were seen for 12/23 NHL. No floor effects (~50%) were seen.

For the 3 adaptive tests (e.g. MedEl MuSIC Test, UW CAMP, SOECIC MTB PDT), defining floor and ceiling effects was complicated by their different inclusion of interval sizes. Floor effects were determined by the poorest possible scores within each test, e.g. the largest interval size tested. For the MedEl MuSIC Test this was 45 semitones, for the UW CAMP test, this was 12 semitones, and for the SOECIC MTB PDT, this was 16 semitones. Ceiling effects were determined by the smallest possible interval within each test. For the MedEl MuSIC Test, this was 0.5 semitone, the UW CAMP it was 1 semitone and for the SOECIC MTB PDT it was 0.01 semitone.

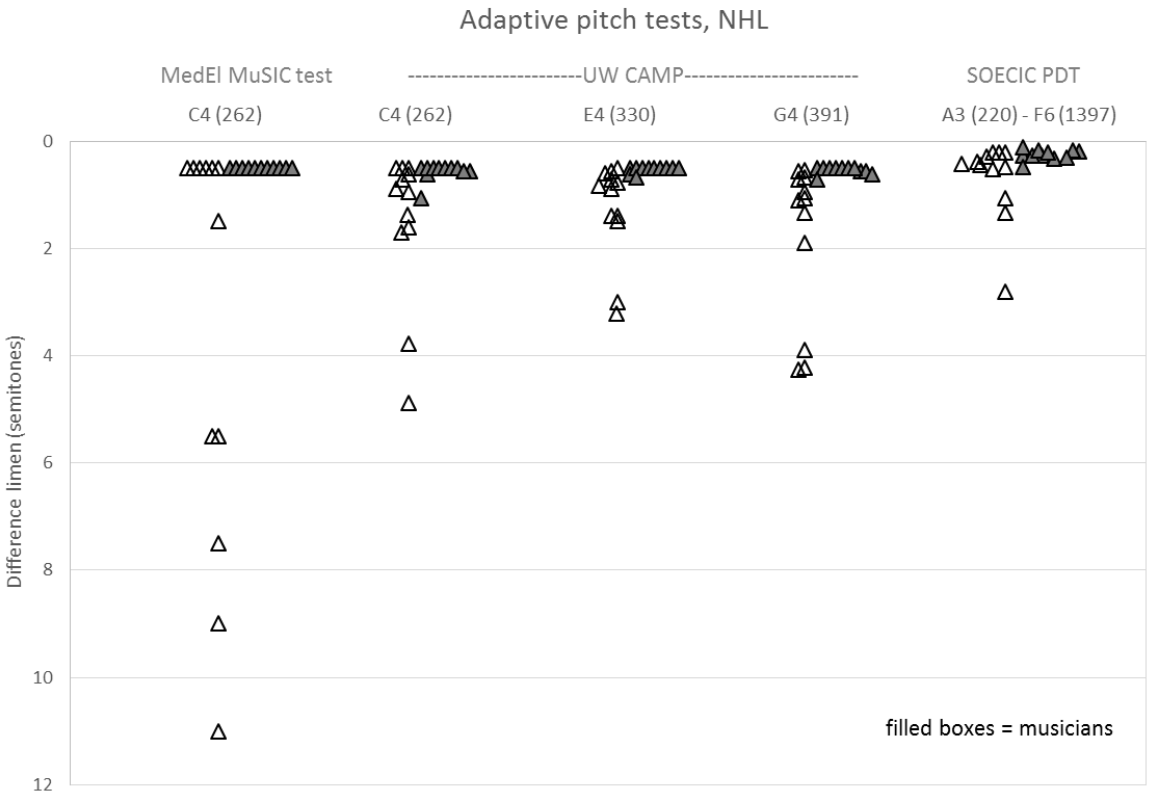


Figure 5.6: Adaptive tests: MedEl MuSIC Test, UW CAMP and SOECIC MTB PDT scores with NHL, T1 data.

MedEl MuSIC Test: ceiling effects were seen for 17/23 NHL. No floor effects were seen.

UW CAMP: ceiling effects were seen for 17/23 NHL (262 Hz), 18/23 NHL (330 Hz), 16/23 NHL (392 Hz). No floor effects were seen.

SOECIC MTB PDT: no ceiling or floor effects were seen.

5.4.3 Test comparisons

It was hypothesised that significantly differing scores would be achieved as a result of test choice, when using the same sample of participants. However, given the ceiling effects seen in this group, it was not considered appropriate to perform any statistical investigation. Generally, the NHL group were performing at ceiling and close to 100% for the 3 MCS tests, and at ceiling for the MedEl MuSIC Test and UW CAMP tests, and achieving scores of < 0.5 semitone on the SOECIC MTB PDT. It was interesting however, to look at two individual NHL, both who performed at a poorer

level than the majority of the group. NHL 4 and NHL 24 achieved the poorest scores in the MedEl MuSIC Test, and generally achieved much poorer scores for the other 5 tests as well.

Table 5.5 The poorest NHL performers

test	NHL 4	NHL 24	NHL medians
PMMA	77.5%	97.5%	100%
MACarena	84%	78.1%	100%
MBEA	80.7%	90.3%	90.32%
MedEl MuSIC Test	9 semitones	11 semitones	0.5 semitones
UW CAMP 262 Hz	1.61 semitones	4.89 semitones	1 semitone
UW CAMP 330 Hz	1.5 semitones	0.56 (1) semitones	1 semitone
UW CAMP 392 Hz	3.89 semitones	4.22 semitones	1 semitone
SOECIC MTB PDT	1.06 semitones	1.33 semitones	0.29 semitone

5.4.4 Sensitivity to musicianship

The sensitivity of each test to a known difference in ability, musicianship, was investigated. There were 11 musicians and 12 non-musicians within the NHL group. Mean rank scores for non-musicians were consistently poorer when compared to musicians, which was significant for all tests except the MBEA (Table 5.6).

Table 5.6 NHL musician and non-musician comparison

Mann-Whitney U test. Musicians (n=11), non-musician (n=12) UW CAMP scores adjusted to 1 for < 1. *U* represents the test statistic, *z* represents where the test statistic falls in relation to the normal distribution, *p* represents the probability of that result occurring due to chance and *r* represents the effect size.

		Non-musician median	*Musician median*	*U*	*z*	*p*	*r*
(%)	MACarena*	97.92	100	38.5	-2.08	0.03	-0.43
	MBEA	88.71	93.55	43	-1.43	0.08	-0.30
	PMMA*	96.25	100	38.5	-1.83	0.04	-0.38
(st)	MedEl MuSIC Test*	1	0.5	33	-2.63	0.01	-0.55
	UW CAMP 262 Hz*	1	1	42	-1.91	0.02	-0.40
	UW CAMP 330 Hz*	1	1	38.5	-2.35	0.02	-0.49
	UW CAMP 392 Hz*	1.09	1	27.5	-2.91	0.00	-0.61
	SOECIC MTB PDT*	0.44	0.27	26	-2.47	0.01	-0.52

**Significant at the p< 0.05 level, one-tailed*

5.4.5 Summary

In summary, both the good and poor performer examples of the UW CAMP (Figure 5.3 and Figure 5.4) were shown to have a sufficient number of successful trials in order to keep $p < 0.05$. This was also true for the poor performer example in the MedEl MuSIC Test (Figure 5.2), however the good performer example (Figure 5.1) was shown to have insufficient trials.

All 3 MCS tests showed ceiling effects with NHL. Both the UW CAMP and the MedEl MuSIC Test also showed ceiling effects with NHL, the SOECIC MTB PDT was the only test that did not show any ceiling effects with NHL.

No statistical comparisons were attempted due to ceiling effects. All tests except for the MBEA showed sensitivity to musicianship status.

5.5 NHL Reliability with SOECIC MTB PDT

Time constraints and ceiling effects made the possibility of assessing test-retest validity using NHL inappropriate. The SOECIC MTB PDT was not affected by ceiling effects and so this test was used in Study 2 in order to determine test-retest reliability. Eighteen NHL completed the SOECIC MTB PDT twice. Data distribution was found to be significantly non-normal, using the Shapiro-Wilk test. No significant difference was seen between T1 (median = 0.21 semitones) and T2 (median = 0.25 semitones), $T = 82$, $p = 0.9$, $r = -0.025$. For comparison, the median score from Study 1 (N = 15) for SOECIC MTB PDT was 0.29 semitone.

Test retest reliability was assessed using the ICC (A,1, Shrout and Fleiss, 1979) and the critical value of *r* for *n* was used, e.g. the amount of correlation that might be expected due to chance.

This meant that the reliability criteria for this thesis was determined by:

1. A coefficient of > 0.8
2. A coefficient significantly greater than the critical value of *r* for *n*, using Pearson's table of critical values, in order to assess whether the ICC was likely to have occurred due to chance alone

A moderately high level of reliability was seen for the SOECIC MTB PDT with NHL. An ICC (A, 1) of 0.75 was seen with a 95% confidence interval ranging from 0.45 - 0.90, and was considered significantly different from a chance level of 0.468 (the two-tailed critical value of *r* for $n = 18$, degrees of freedom (df) = n-2 = 16; $F(17,17) = 2.534$, $p = 0.032$). A graphical display of the T1 and T2 data can be seen in Figure 5.7.

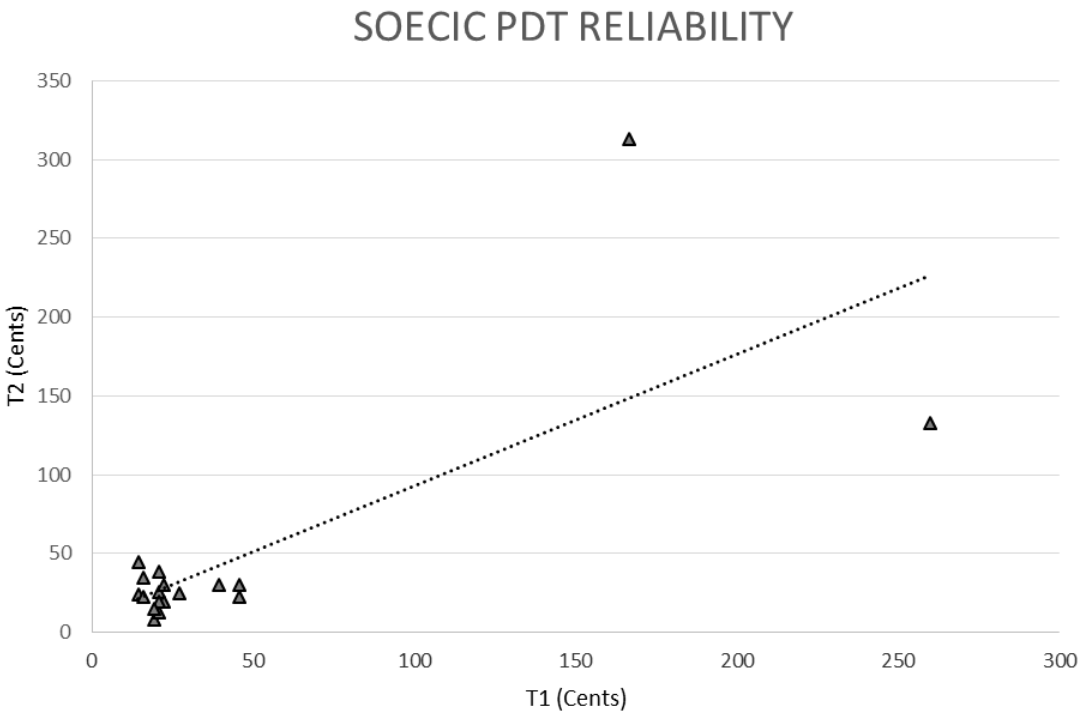


Figure 5.7: SOECIC MTB PDT reliability data with NHL, T1 & T2, n = 15

Two data points were considered to be outliers compared to the rest of the NHL data. The two NHL that were only able to achieve very large pitch discrimination thresholds using the SOECIC MTB PDT struggled with this test both at T1 and T2, and the researcher considered that they may have some form of amusia and were not aware of it themselves. The inclusion of these two data points in this reliability analysis meant that the ICC appeared to indicate a much greater reliability than may have been the case. As such, the reliability analysis was repeated with these two data points removed, and the updated data set is displayed in Figure 5.8.

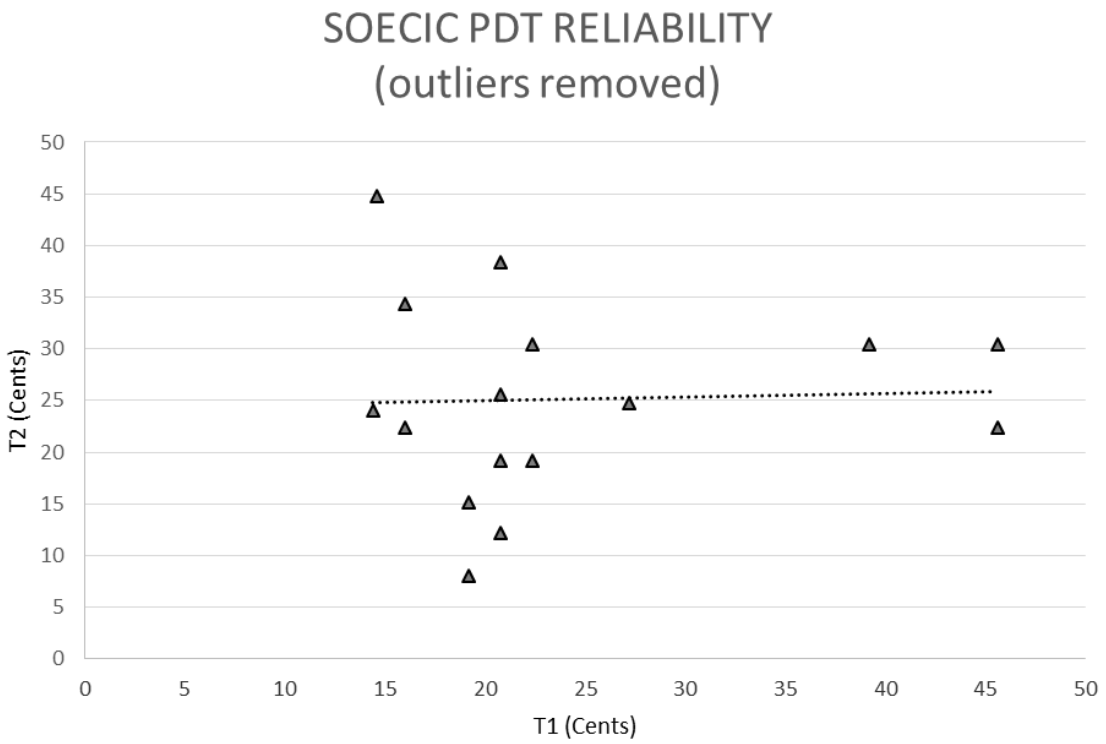


Figure 5.8: SOECIC MTB PDT reliability data with NHL, T1 & T2, with suspected outliers (NHL 8 & 16) removed, n = 13

With the two outliers removed, the reliability decreased, ICC (A, 1) = 0.039, with a 95% confidence interval ranging from -0.49 - 0.53, and was no longer considered significantly different from the chance level of 0.497 (for n = 16; df = 14, F (15,15) = 0.376, p = 0.966).

5.5.1 Summary

Initially, the SOECIC MTB PDT appeared highly reliable with NHL. However with the removal of two outliers (due to possible amusia), reliability coefficients dropped, indicating a less than desirable reliability for this test, with the NHL population.

5.6 CI Results

This section presents the results of the PMMA, the MedEl MuSIC Test, the UW CAMP, the SOECIC MTB PDT, and the MCI using a group of 15 CI users. Results from T1 only were analysed. Data from the UW CAMP was adjusted so that any scores of < 1 were reassigned the value of 1, except for when testing reliability on retest.

Each test was analysed to determine whether the number of trials was sufficient to minimise the effects of chance on results; whether floor and ceiling effects affected the data; how each test fared in terms of test-retest reliability, comparisons were made between the tests; and sensitivity to musicianship was investigated.

Data distribution was found to be normally distributed except for UW CAMP E4 (330 Hz) and MCI 3 semitones. As such, both means, medians, SD and IQR are presented below, and non-parametric statistics were used for analysis.

Table 5.7 CI median, IQR, mean, SD, maximum and minimum scores for all tests (UW CAMP data set to 1 if < 1)

test	time	median	IQR	mean	SD	min	max
PMMA (%)	T1	72.5	15	69.17	11.56	42.5	90
	T2	70	11.25	69.83	9.70	52.5	92.5
MedEl MuSIC Test	T1	15.75	6.13	15.29	6.41	2	27
(semitones)	T2	16.75	17.5	15.07	9.79	0.5	27
UW CAMP 262 Hz	T1	3.06	3.67	4.59	3.18	1	11.94
(semitones)	T2	3.33	2.11	3.82	2.72	1	9.56
UW CAMP 330 Hz	T1	1.56	1.64	2.51	2.17	1	6.61
(semitones)	T2	1.61	1.84	2.79	2.54	1	9.22
UW CAMP 392 Hz	T1	2.72	3.58	3.30	2.31	1	7.61
(semitones)	T2	1.39	1.84	2.56	2.29	1	7.83
SOECIC MTB PDT	T1	2.6	2.14	2.81	1.45	0.62	5.4
(semitones)	T2	2.6	3.1	3.6	2.44	0.22	9.2
MCI 5 (%)	T1	66.67	29.65	64.19	23.66	22.22	92.59
	T2	74.08	35.19	73.55	23.91	22.22	100
MCI 4 (%)	T1	62.96	33.34	63.27	27.14	14.81	96.59
	T2	74.07	29.63	70.09	23.93	22.22	96.3
MCI 3 (%)	T1	74.07	44.45	62.39	26.66	22.22	92.59
	T2	55.56	48.15	64.96	28.95	18.52	100
MCI 2 (%)	T1	62.96	48.15	57.27	29.42	7.41	100
	T2	55.56	40.75	58.98	27.78	18.52	96.3
MCI 1 (%)	T1	48.15	37.04	45.87	28.71	3.7	96.3
	T2	33.33	40.74	45.54	31.89	3.7	100

5.6.1 Trial number and the role of chance in adaptive methods

As described in section 5.4.1 above, the binomial calculator was used to determine the likelihood of final scores occurring due to chance, for the adaptive tests.

MedEl MuSIC Test: 2 AFC, chance = 50%, 3 down 1 up, no repeats

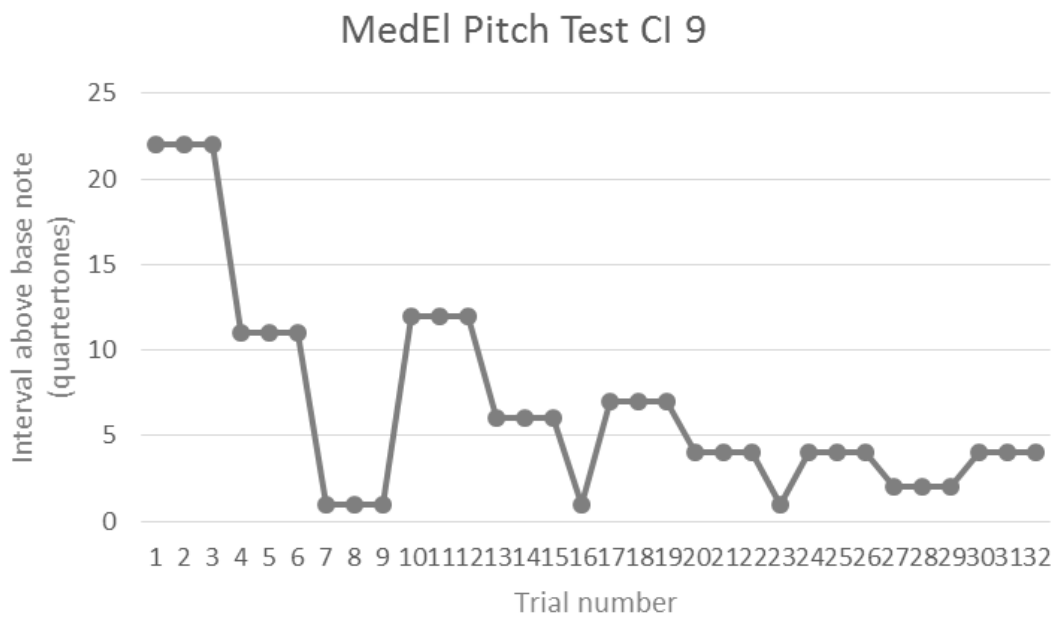


Figure 5.9 Example of a good performer (CI 9) using the MedEl MuSIC Test Pitch Test, showing the adaptive staircase. Final score was 4 quartertones (2 semitones) and this test (with this level of performance) had sufficient trials to ensure that the final score had $p < 0.05$.

In the example above, the final score was 4 quartertones (2 semitones). The final score of the MedEl MuSIC Test is approximated by averaging the final 5 of the 8 reversals. Figure 5.9 shows the final 5 reversals (in reverse order) as 4, 2, 4, 1 and 7, and their average is 3.6 quartertones, a close approximation to the final score obtained by the test, of 4 quartertones.

Below is a breakdown of CI 9's performance, for each interval presented, the number of successes and the number of trials. The MedEl MuSIC Test used intervals measured in quartertones.

22 quartertones: 3/3 successful

12 quartertones: 3/3 successful

11 quartertones: 3/3 successful

7 quartertones: 3/3 successful

6 quartertones: 3/3 successful

4 quartertones: 9/9 successful*

2 quartertones: 2/3 successful

1 quartertone: 2/5 successful

As shown in section 5.3.4, when the level of chance is 50%, 5 trials (or more) are needed to keep $p < 0.05$, and these are indicated by *. This indicates that enough successful trials were presented for 4 quartertones. The final score was 4 quartertones, and so the likelihood that this was due to chance is small.

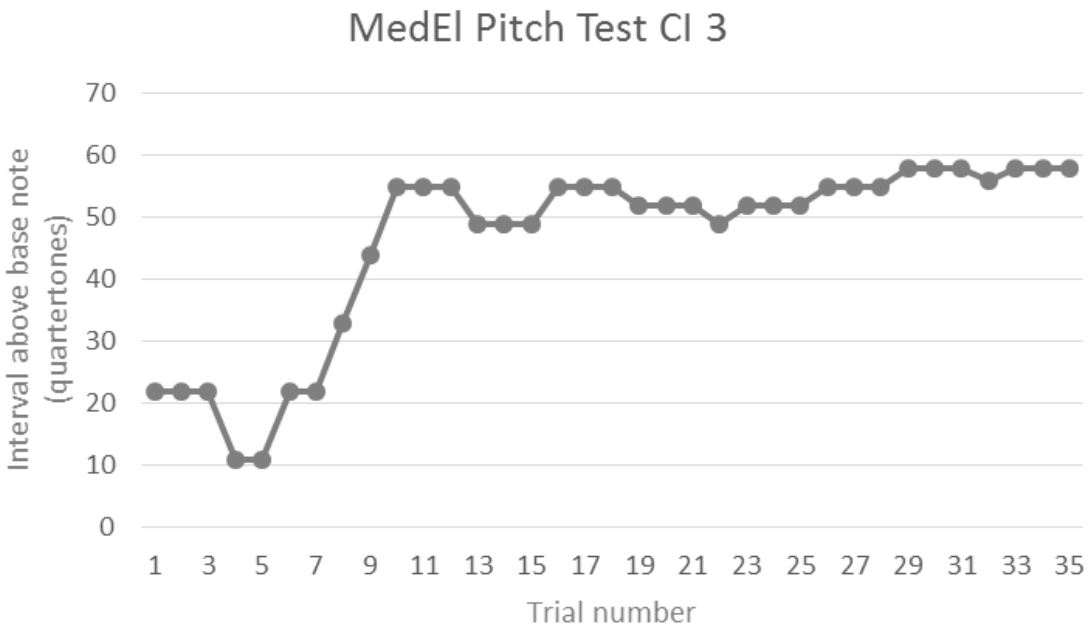


Figure 5.10 Example of a poor performer (CI 3) using the MedEl MuSIC Test Pitch Test, showing the adaptive staircase. Final score was 54 quartertones (27 semitones) and this test (with this level of performance) had sufficient trials to ensure that the final score had $p < 0.05$.

In the example above, the final score was 54 quartertones (27 semitones). Figure 5.10 shows the final 5 reversals (in reverse order) as 58, 56, 58, 49 and 55, and their average is 55.2 quartertones, a close approximation to the final score obtained by the test, of 54 quartertones.

Below is a breakdown of CI 3's performance, for each interval presented, the number of successes and the number of trials.

11 quartertones: 1/2 successful

22 quartertones: 4/5 successful

33 quartertones: 0/1 successful

44 quartertones: 0/1 successful

49 quartertones: 2/4 successful

52 quartertones: 5/6 successful

55 quartertones: 8/9 successful*

56 quartertones: 0/1 successful

58 quartertones: 6/6 successful*

This indicates that enough successful trials were presented for 55 and 58 quartertones in order to be sure that the likelihood that this was due to chance was less than 5%. The final score was 54 quartertones, which was an interval that was not tested, however it is close to the interval 55 quartertones, and so the likelihood that this was due to chance is small.

UW CAMP Pitch Test: 2 AFC, chance = 50%, 1 down 1 up, 3 repeats

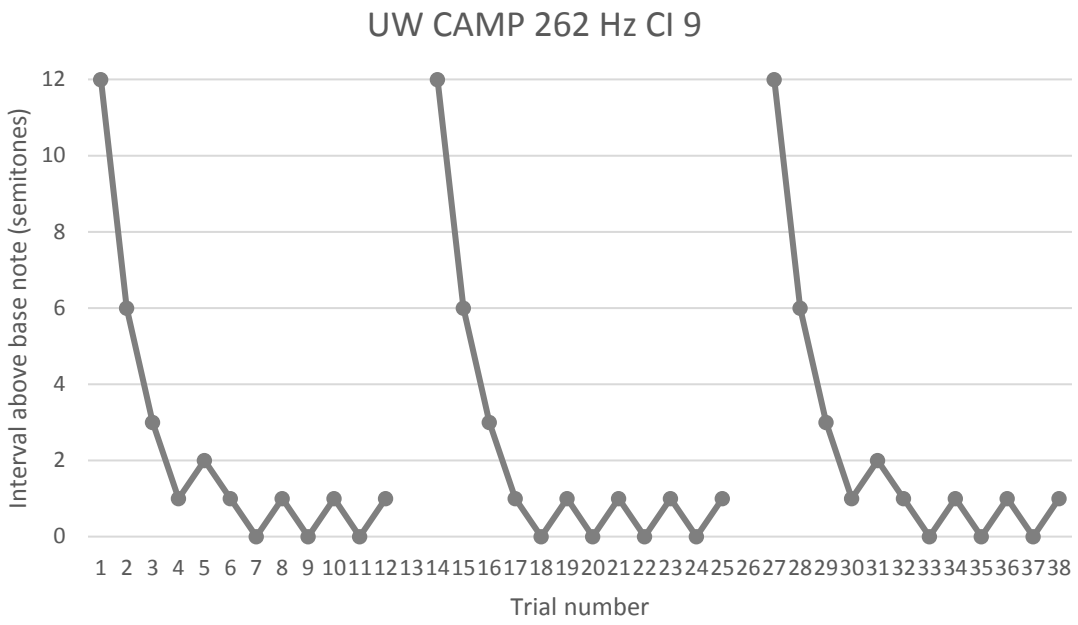


Figure 5.11 Example of a good performer (CI 9) using the UW CAMP Test (262 Hz), showing the adaptive staircase. Final score was 0.5 semitone (run 1: 0.5, run 2: 0.5, run 3: 0.5) and this test (with this level of performance and three repeats) had sufficient trials to ensure that the final score had $p < 0.05$.

In the example above, the final score was 0.5 (1) semitones. Although runs 1 and 3 show an initial mistake at 1 semitone, because the final 6 reversals are averaged, all 3 runs have the same final 6 reversals of 1, 0, 1, 0, 1 and 0, as can be seen in Figure 5.11. The average of the 3 runs is therefore 0.5 semitone.

Below is a breakdown of CI 9's performance, for each interval presented, the number of successes and the number of trials.

12 semitones: 3/3 successful

6 semitones: 3/3 successful

3 semitones: 3/3 successful

2 semitones: 2/2 successful

1 semitone: 13/15 successful* (cumulative binomial probability of 13/15 = 0.004)

This indicates that enough successful trials were presented for 1 semitone. The final score was 1 semitone, and so the likelihood that this was due to chance is small.

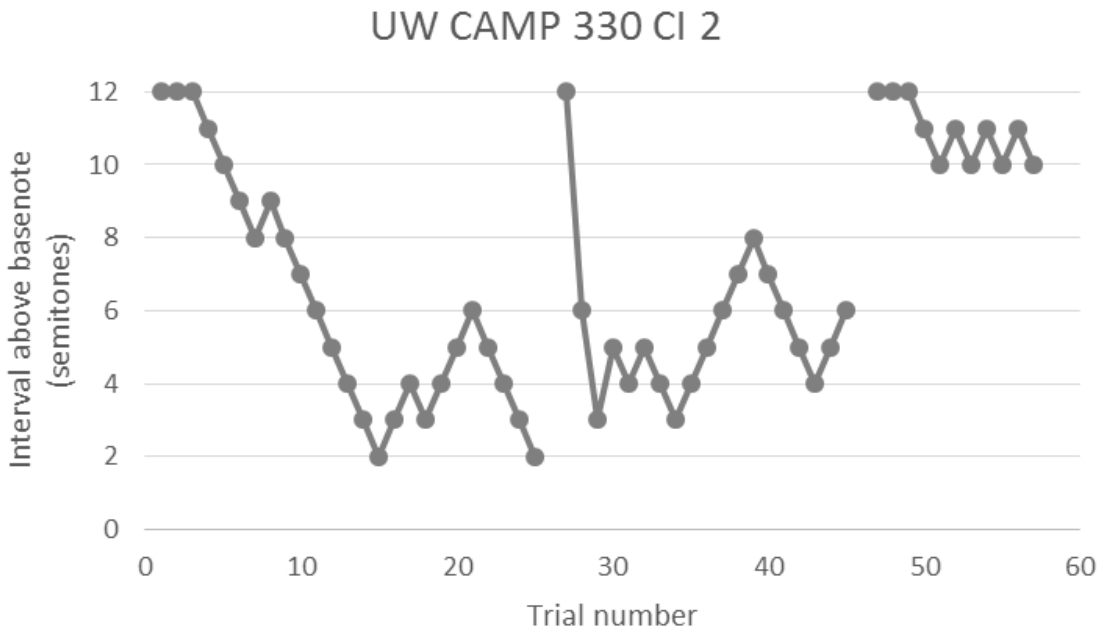


Figure 5.12 Example of a poor performer (CI 2) using the UW CAMP Test (330 Hz), showing the adaptive staircase. Final score was 6.61 semitones (run 1: 4.33, run 2: 5, run 3: 10.5) and this test (with this level of performance and three repeats) had insufficient trials to ensure that the final score had $p < 0.05$.

In the example above, the final score was 6.61 semitones. Figure 5.12 run 1 shows the final 6 reversals (in reverse order) as 2, 6, 3, 4, 2 and 9, and their average is 4.33. Run 2 shows the final 6 reversals as 6, 4, 8, 3, 5 and 4 and their average is 5. Run 3 shows the final 6 reversals as 10, 11, 10, 11, 10 and 11 and their average is 10.5. The average of these 3 run final scores is 6.61 semitones.

Below is a breakdown of CI 2's performance, for each interval presented, the number of successes and the number of trials.

12 semitones: 7/7 successful*

11 semitones: 5/5 successful*

10 semitones: 1/5 successful

9 semitones: 2/2 successful

8 semitones: 2/3 successful

7 semitones: 2/3 successful

6 semitones: 5/6 successful

5 semitones: 5/8 successful

4 semitones: 4/8 successful

3 semitones: 2/6 successful

2 semitones: 0/2 successful

This indicates that there were not enough successful trials presented for any of the intervals except 12 and 11 semitones. The final score was 6.61 semitones, and due to the low number of successful trials at 6 semitones, the likelihood that this was due to chance is high.

SOECIC MTB PDT: 3 AFC, chance = 33%, 2 down 1 up, no repeats

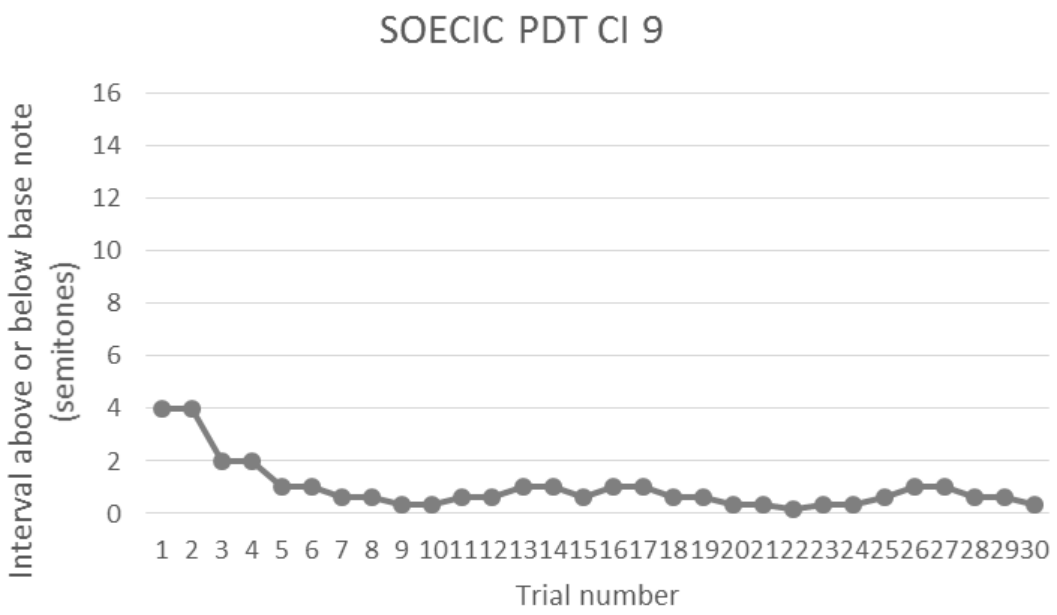


Figure 5.13 Example of a good performer (CI 9) using the SOECIC MTB PDT, showing the adaptive staircase. Final score was 0.62 semitone and this test (with this level of performance) had sufficient trials to ensure that the final score had $p < 0.05$.

In the example above, the final score was 0.62 semitones. Figure 5.13 shows the final 5 reversals (in reverse order) as 0.32, 1, 0.16, 1 and 0.64 semitone, and their average is 0.62 semitone.

Below is a breakdown of CI 9's performance, for each interval presented, the number of successes and the number of trials.

4 semitones:	2/2 successful
2 semitones:	2/2 successful
1 semitone:	8/8 successful*
0.64 semitone:	7/10 successful*
0.32 semitone:	4/7 successful
0.16 semitone:	0/1 successful

This indicates that enough successful trials were presented for 0.64 semitones. The final score was 0.62 semitones, and so the likelihood that this was due to chance is small.

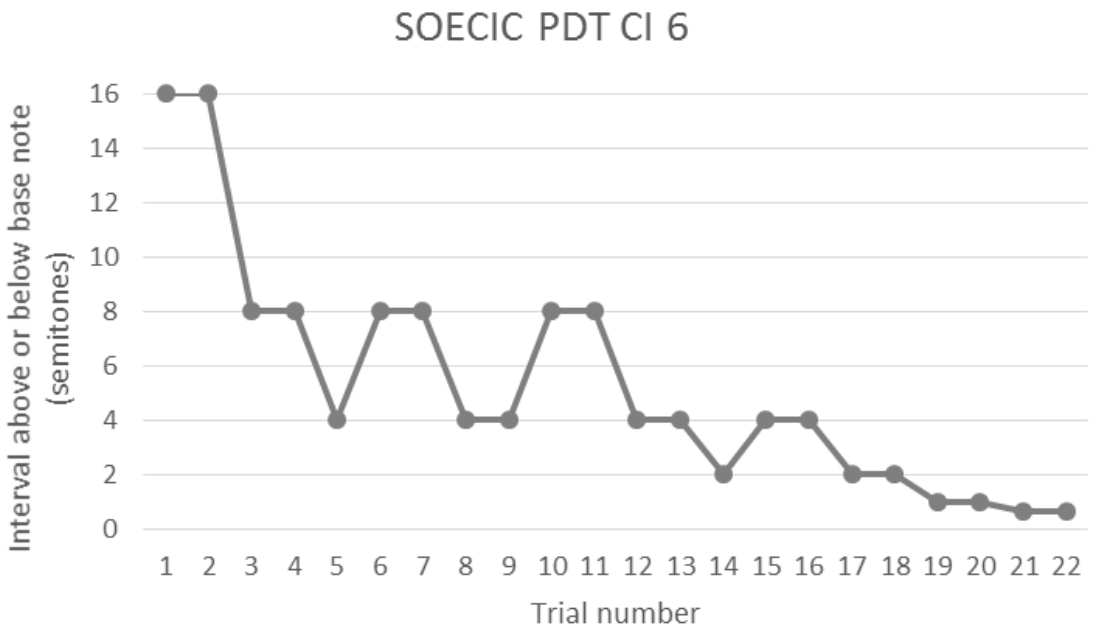


Figure 5.14 Example of a poor performer (CI 6) using the SOECIC MTB PDT, showing the adaptive staircase. Final score was 3.73 semitones and this test (with this level of performance) had sufficient trials to ensure that the final score had $p < 0.05$.

In the example above, the final score was 3.73 semitones. Figure 5.14 shows the final 5 reversals (in reverse order) as 0.64, 4, 2, 8 and 4, and their average is 3.73 semitones.

Below is a breakdown of CI 6's performance, for each interval presented, the number of successes and the number of trials.

16 semitones: 2/2 successful

8 semitones: 6/6 successful*

4 semitones: 5/7 successful* (cumulative binomial probability, chance = 0.33, of 5/7 = 0.045)

2 semitones: 2/3 successful

1 semitone: 2/2 successful

0.64 semitone: 2/2 successful

This indicates that enough successful trials were presented for 8 and 4 semitones. The final score was 3.73 semitones, and so the likelihood that this was due to chance is small.

5.6.2 Floor and ceiling effects with CI

For the MCS tests (e.g. PMMA and MCI), ceiling effects were defined as scores of 100%, and scores at chance level (50% for PMMA and 11% for the MCI) were considered to be floor effects.

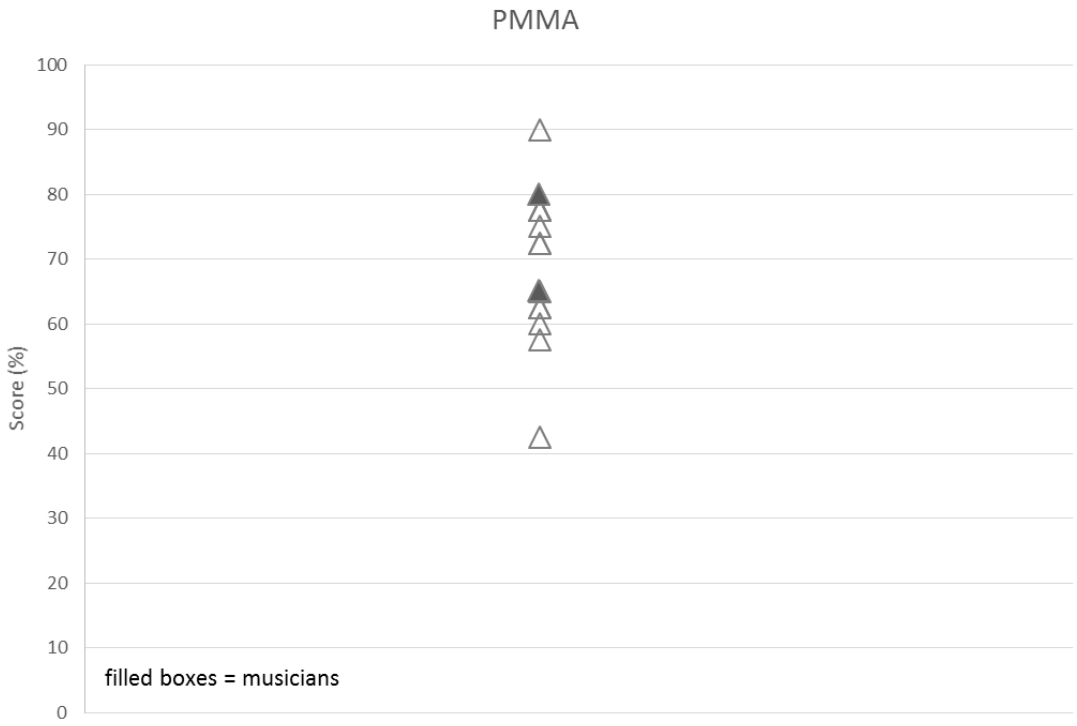


Figure 5.15: PMMA scores with CI users, T1 data

PMMA: ceiling effects (100%) were not seen. Floor effects were seen for 1/15 CI (CI 3).

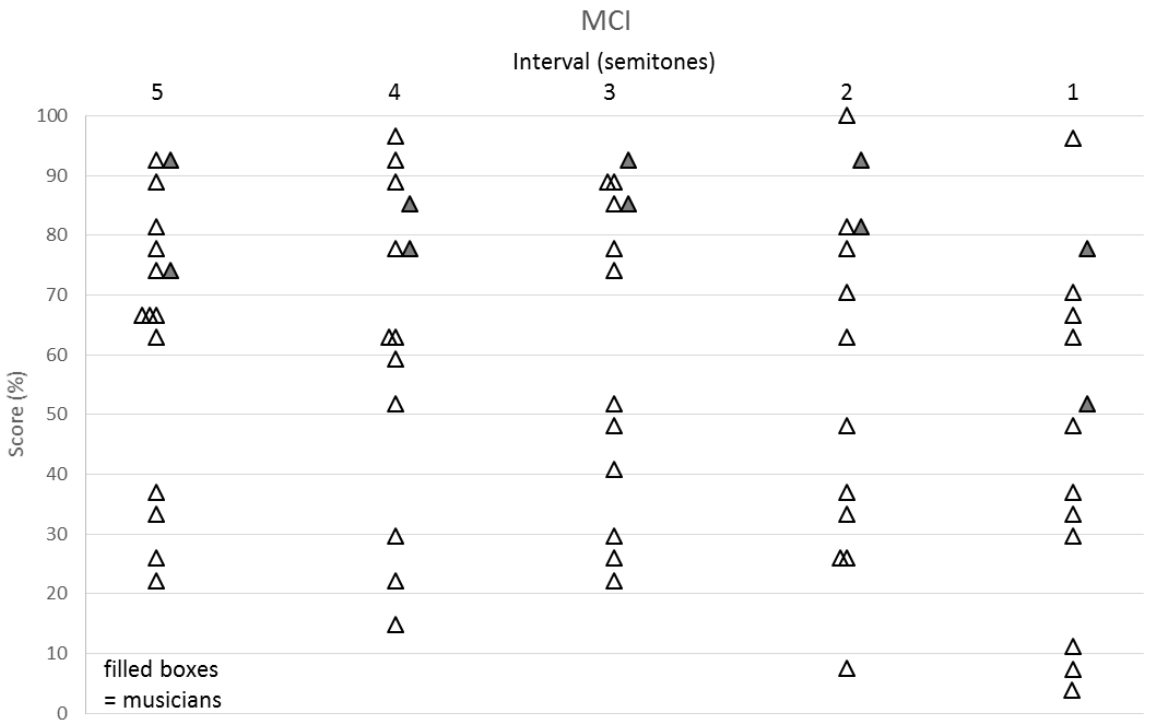


Figure 5.16: MCI scores with CI users, T1 data

MCI 5: no ceiling or floor effects were seen.

MCI 4: no ceiling effects were seen. Floor effects were seen for 1/15 CI (CI 5).

MCI 3: no ceiling or floor effects were seen.

MCI 2: ceiling effects (100%) were seen for 1/15 CI (CI 9). Floor effects were seen for 1/15 CI (CI 15).

MCI 1: no ceiling effects were seen. Floor effects were seen for 3/15 CI (CI 5, CI 10, CI 15).

For the adaptive tests (e.g. MedEl MuSIC Test, UW CAMP, SOECIC MTB PDT), defining floor and ceiling effects was complicated by their different inclusion of interval sizes. Floor effects were determined by scores at the top end of the largest interval, so for the MedEl MuSIC Test this was 45 semitones, the UW CAMP test, this was 12 semitones (scores >11 semitones were considered to be floor), and for the SOECIC MTB PDT, this was 16 semitones. Ceiling effects were determined by the smallest possible interval within each test. For the MedEl MuSIC Test this was 0.5 semitone, the UW CAMP it was 1 semitone and for the SOECIC MTB PDT it was 0.01 semitone.

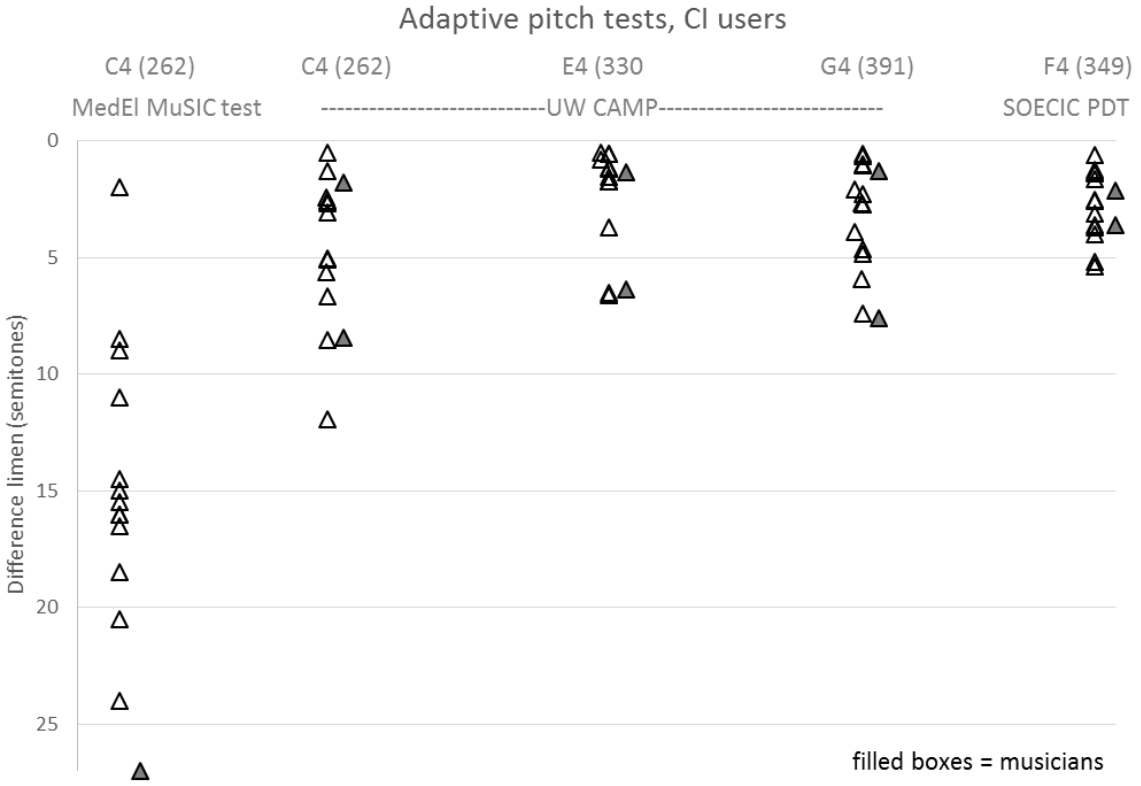


Figure 5.17: Adaptive pitch tests: MedEl MuSIC Test, UW CAMP and SOECIC MTB PDT test scores with CI users, T1 data. Please note the differing base notes of the SOECIC MTB PDT compared to NHL Figure 5.6.

MedEl MuSIC Test: no ceiling effects were seen. Floor effects were seen for 1/15 CI: CI 1. These results will be presented in more detail below.

UW CAMP 262 Hz: ceiling effects were seen for 1/15 CI (CI 9). Floor effects were seen for 1/15 CI (CI 2).

UW CAMP 330 Hz: ceiling effects were seen for 4/15 CI (CI 4, CI 5, CI 6, CI 9). No floor effects were seen.

UW CAMP 392 Hz: ceiling effects were seen for 3/15 CI (CI 2, CI 4, CI 9). No floor effects were seen.

SOECIC MTB PDT: no ceiling or floor effects were seen.

MedEl MuSIC Test floor effect

CI 1 was scored '0' quartertones at T1 and T2 on the MedEl MuSIC Test, due to failing to get 3 successful responses at 96 quartertones.

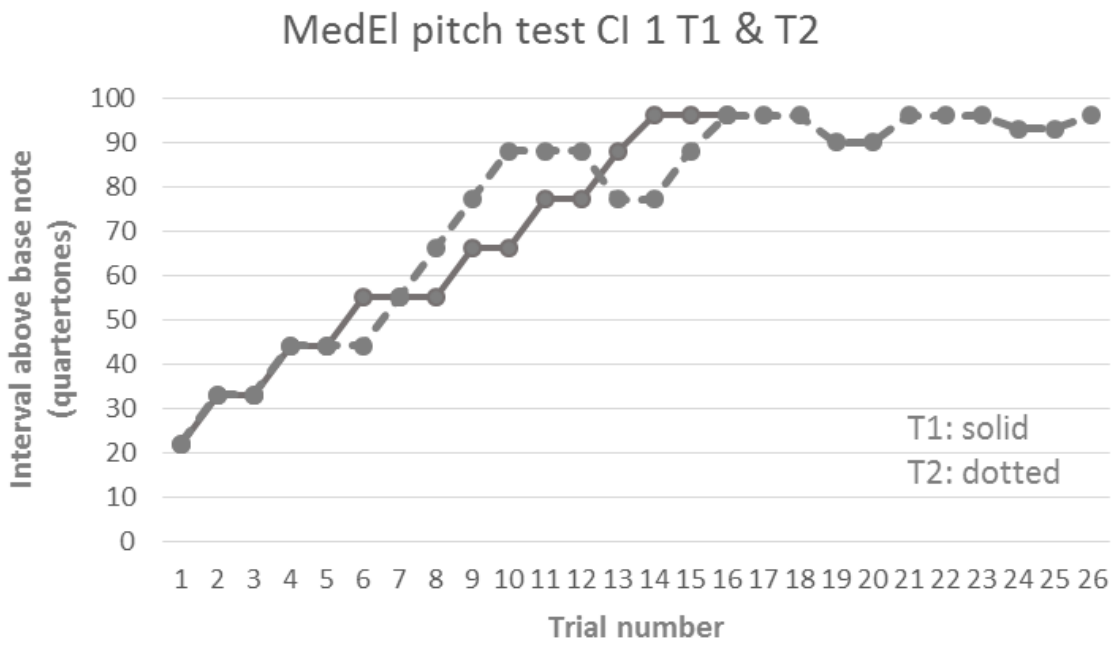


Figure 5.18: MedEl MuSIC Test scores for CI 1 at T1 and T2, showing their similarity and their termination at the largest interval of 96 quartertones (48 semitones). The final score for both of these runs was '0' quartertones.

The T1 test ended after 3 presentations at 96 quartertones: correct, correct, incorrect, test ended. The T2 test ended when the first presentation at 96 quartertones was incorrect as it could not increase the interval further. Then test's output did not state anything other than final score: 0 quartertones. As a result, CI 1's data had to be excluded from any further analysis.

5.6.3 CI Reliability

Fifteen CI users took part in each test twice, allowing test retest reliability of the 5 pitch tests to be assessed. Test retest reliability for the PMMA, MedEl MuSIC Test, UW CAMP, SOECIC MTB PDT and MCI scores was assessed using the ICC (A,1, Shrout and Fleiss, 1979). The UW CAMP data were kept in their original format, e.g. if they were < 1 semitone, this was kept so that the reliability data would be more accurate in comparing T1 and T2. Reasons for choosing the ICC over the Pearson's *r* are described in section 5.5. Reliability criteria for tests used with CI users was determined by:

1. A coefficient of > 0.8
2. A coefficient significantly greater than the critical value of *r* for *n*.

This criteria was met for the 262 Hz and 330 Hz of the UW CAMP (Figure 5.19 and Figure 5.20), and all the MCI tests (Figure 5.21, Figure 5.22, Figure 5.23, Figure 5.24 and Figure 5.25).

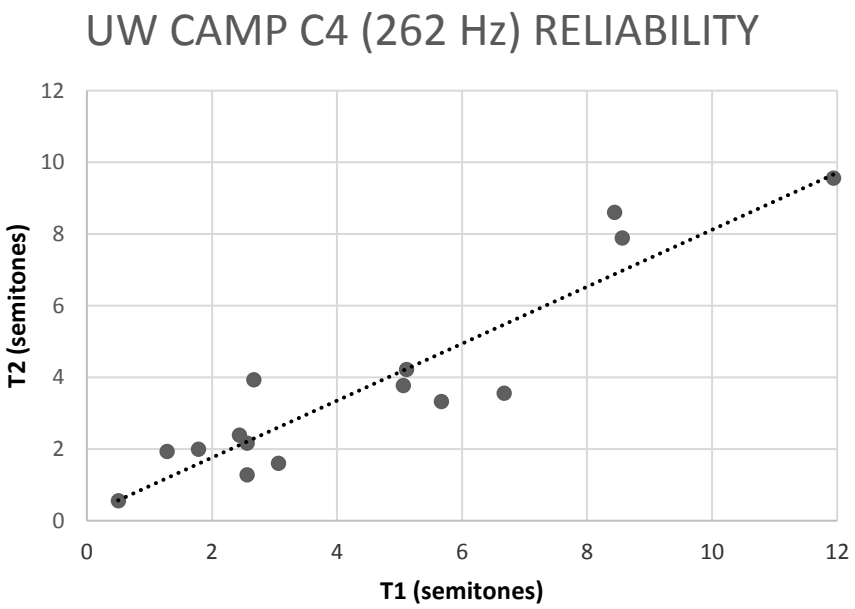


Figure 5.19 UW CAMP pitch test (C4, 262 Hz) reliability data with CI users, *n* = 15. Data from T1 and T2 met the criteria of an ICC of > 0.8 and were significantly different from the critical value of *r* for *n* (*df* = *n*-2 = 13, *r* = 0.44)

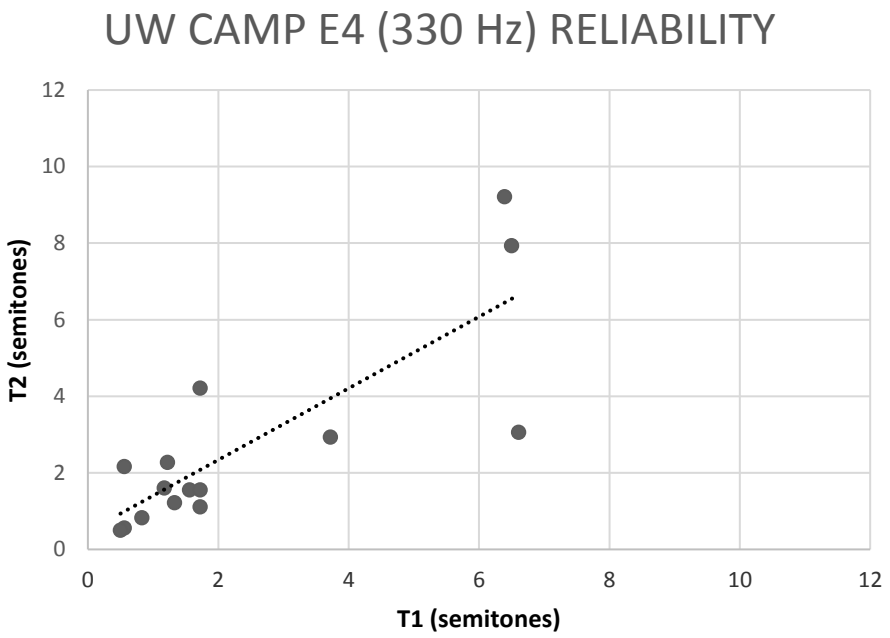


Figure 5.20 UW CAMP pitch test (E4, 330 Hz) reliability data with CI users, *n* = 15. Data from T1 and T2 met the criteria of an ICC of > 0.8 and were significantly different from the critical value of *r* for *n* (*df* = *n*-2 = 13, *r* = 0.44)

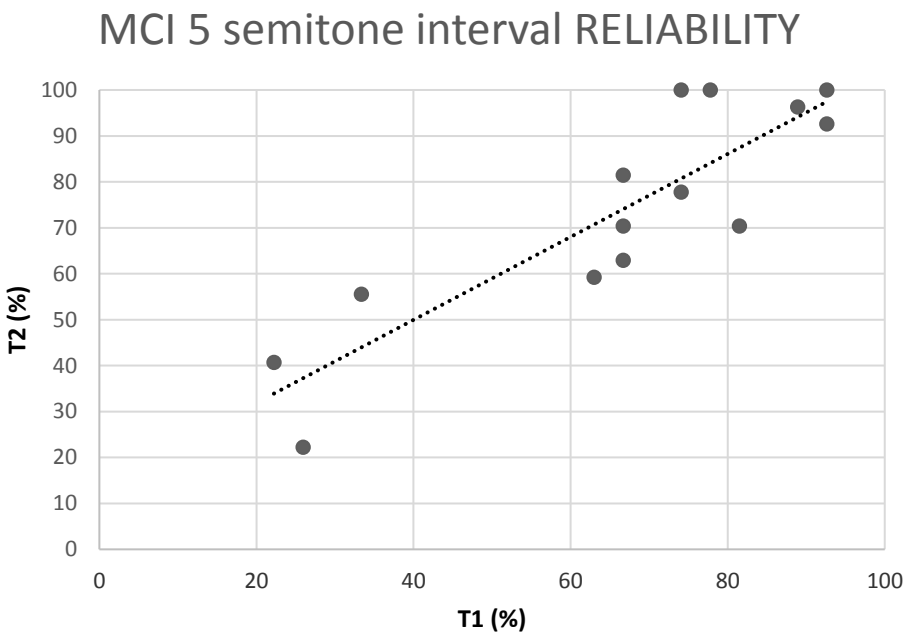


Figure 5.21 MCI test (5 semitones) reliability data with CI users, *n* = 14. Data from T1 and T2 met the criteria of an ICC of > 0.8 and were significantly different from the critical value of *r* for *n* (*df* = *n*-2 = 12, *r* = 0.46)

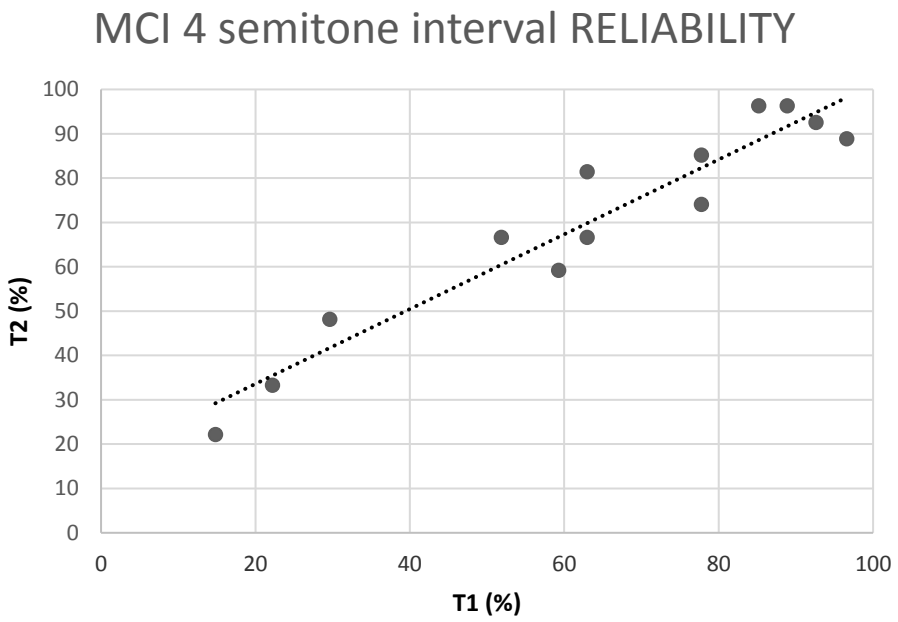


Figure 5.22 MCI test (4 semitones) reliability data with CI users, *n* = 13. Data from T1 and T2 met the criteria of an ICC of > 0.8 and were significantly different from the critical value of *r* for *n* (*df* = *n*-2 = 11, *r* = 0.48)

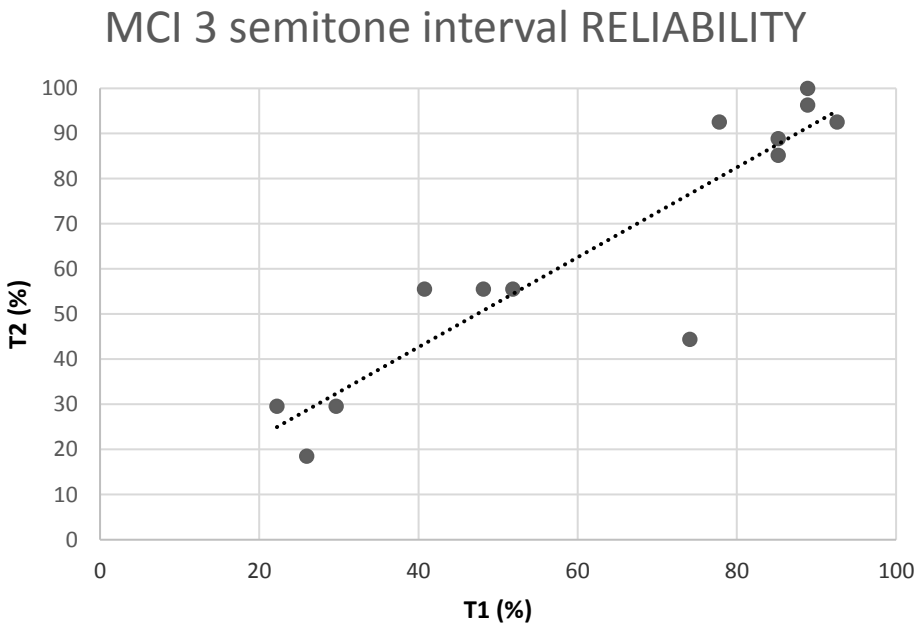


Figure 5.23 MCI test (3 semitones) reliability data with CI users, *n* = 13. Data from T1 and T2 met the criteria of an ICC of > 0.8 and were significantly different from the critical value of *r* for *n* (*df* = *n*-2 = 11, *r* = 0.48)

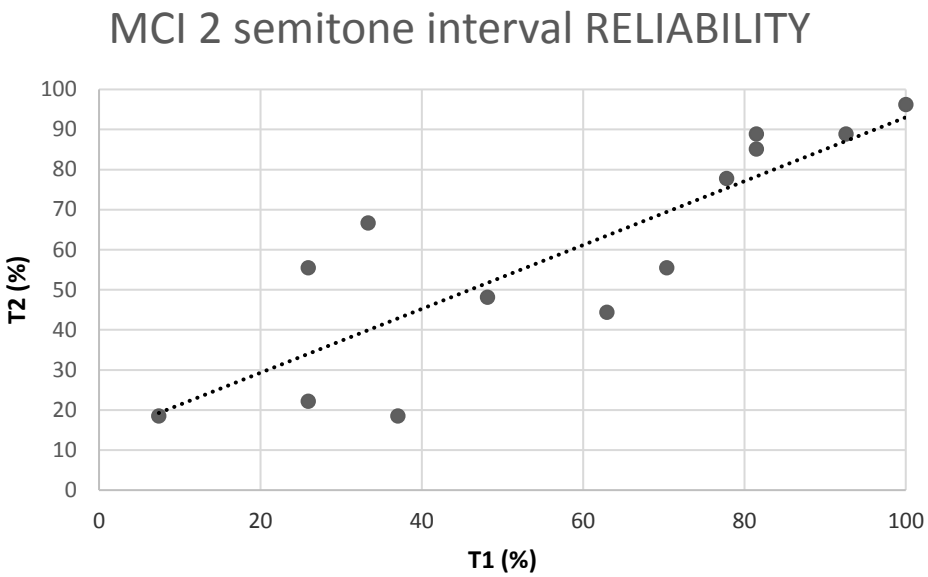


Figure 5.24 MCI test (2 semitones) reliability data with CI users, *n* = 13. Data from T1 and T2 met the criteria of an ICC of > 0.8 and were significantly different from the critical value of *r* for *n* (*df* = *n*-2 = 11, *r* = 0.48)

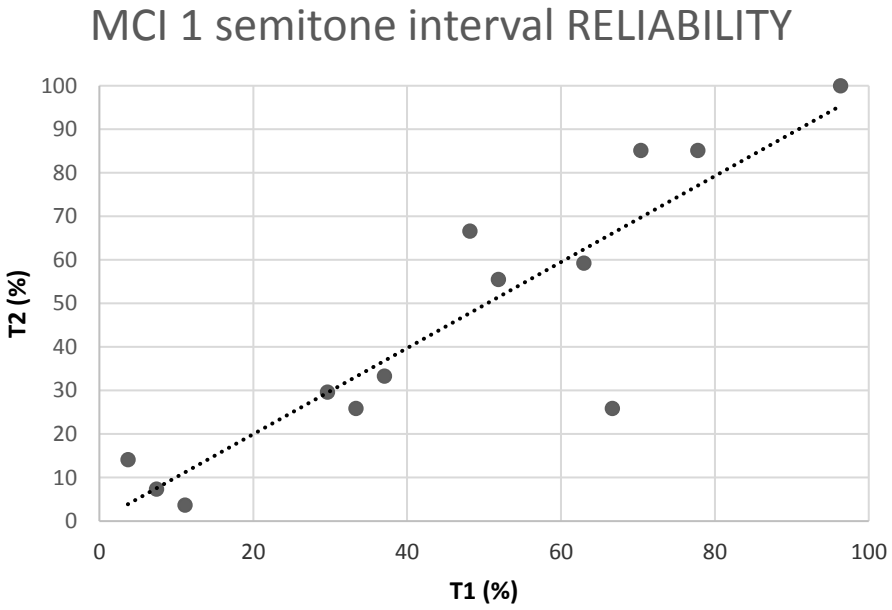


Figure 5.25 MCI test (1 semitone1) reliability data with CI users, *n* = 13. Data from T1 and T2 met the criteria of an ICC of > 0.8 and were significantly different from the critical value of *r* for *n* (*df* = *n*-2 = 11, *r* = 0.48)

Table 5.8 CI user (n=15) Intraclass correlation coefficient

**ICC significantly greater than the critical value for r, at the one tailed p < 0.05 level, and >.80*

Degrees of freedom = n-2. Critical values for r: n=15, r = 0.44; n = 14, r = 0.46; n = 13, r = 0.48

test	n (pairs)	ICC(A,1)	95% ci	F(df1, df2)	sig
PMMA	15	0.77	0.45 – 0.92	3.05(14,14)	0.02
MCI 5*	14	0.84	0.51 – 0.95	4.99 (13,11)	0.01
MCI 4*	15	0.92	0.59 – 0.98	10.76 (12,7)	0.00
MCI 3*	15	0.92	0.76 – 0.97	8.76 (12,13)	0.00
MCI 2*	15	0.85	0.58 – 0.95	4.74 (12,12)	0.01
MCI 1*	15	0.89	0.68 – 0.97	6.65 (12,12)	0.00
MedEl MuSIC Test	14	0.77	0.41 – 0.92	2.91 (13,13)	0.03
UW CAMP 262 Hz*	13	0.89	0.64 – 0.96	7.26 (14,12)	0.00
UW CAMP 330 Hz	13	0.80	0.52 – 0.93	3.53 (14,15)	0.01
UW CAMP 392 Hz	13	0.76	0.41 – 0.91	2.88 (14,14)	0.03
SOECIC MTB PDT	13	0.43	-0.04 – 0.76	0.98 (14,15)	0.51

High levels of reliability, as measured by the ICC (e.g. > 0.75), were seen for all pitch tests except for the SOECIC MTB PDT. All tests except the SOECIC MTB PDT were shown to have reliability coefficients that were significantly greater than the critical value for *r*. All tests except for the SOECIC MTB PDT also showed at least some floor and ceiling effects, which can impact on reliability analysis.

5.6.4 Test comparisons

It was hypothesised that significantly differing scores would be achieved as a result of test choice, when using the same sample of participants, not least because of the large differences in test approach, design and methodologies. It was felt important that these comparisons were made,

however, in order to highlight the differences between tests that may not be discriminable from a clinical point of view. If a clinician wants to test pitch perception, in order to assess the implant's performance, or to address a specific need of the patient, they are likely to use whatever pitch test their department may have available to them. If the patient has been tested in the past (at a different clinic), these test results may even be compared, as little may be known as to the large fundamental differences between the tests. These comparisons were also made to see whether this was likely to be the case or whether because of the overriding shared nature of the tests concept (e.g. to test pitch perception) whether the tests would show similar results regardless of the differences in test design.

Two individual CI users stood out as having particularly interesting results across the tests. CI 9 (latest BKB score of 99%, 2 years implant experience, Advanced Bionics, non-musician) showed overall excellent performance, and CI 15 (pre-lingually deafened, 1 year implanted, MedEl, non-musician) generally showed poor performance on the MCI 4, 2 and 1 but performed very well in the PMMA, see Table 5.9.

Table 5.9 Interesting performer CI users

test	CI 9	CI 15	CI medians
PMMA	90%	80%	72.5%
MedEl MuSIC Test	2 semitones	15.5 semitones	15.75 semitones
UW CAMP 262 Hz	1 semitone	2.44 semitones	3.06 semitone
UW CAMP 330 Hz	1 semitone	1.22 semitones	1.56 semitone
UW CAMP 392 Hz	1 semitone	2.72 semitones	2.72 semitone
SOECIC MTB PDT	0.62 semitone	1.23 semitones	2.6 semitones
MCI 5	92.59%	22.22%	66.67%
MCI 4	92.59%	22.22%	62.96%
MCI 3	88.89%	25.93%	74.07%
MCI 2	100%	7.41%	62.96%
MCI 1	96.3%	3.7%	48.15%

The 3 adaptive tests produced results that could all be measured in semitones and so these results were compared, however the UW CAMP and the MedEl MuSIC Test both tested pitch ranking

ability, whereas the SOECIC MTB PDT tested pitch discrimination. Although the 3 tests were on the whole normally distributed, the UW CAMP 330 Hz scores were non-normally distributed and so non-parametric analysis was chosen to primarily investigate the effects of test on score.

Scores were seen to be significantly affected by the test used to measure them (χ^2 (4) = 35.71, p < .001). Post hoc Wilcoxon analysis, with Bonferroni correction, (Table 5.10) revealed significantly poorer scores for the MedEl MuSIC Test when compared with all other adaptive tests as well as significantly poorer scores for the UW CAMP 262 Hz when compared with 330 Hz.

Table 5.10 CI user comparison of adaptive test scores

Pairwise Wilcoxon signed ranks test (n = 15)

Compared test pairs (median, semitones)			*Mdn*	*T*	*p*	*r*
MedEl MuSIC Test	15.75	UW CAMP 262 Hz*	3.06	0	0.001	-0.38
		UW CAMP 330 Hz*	1.56	0	0.001	-0.38
		UW CAMP 392 Hz*	2.72	0	0.001	-0.38
		SOECIC MTB PDT*	2.6	0	0.001	-0.38
UW CAMP 262 Hz	3.06	UW CAMP 330 Hz*	1.56	0	0.001	-0.38
		UW CAMP 392 Hz	2.72	45	0.394	-0.10
		SOECIC MTB PDT	2.6	27	0.061	-0.22
UW CAMP 330 Hz	1.56	UW CAMP 392 Hz	2.72	21	0.027	-0.26
		SOECIC MTB PDT	2.6	46	0.427	-0.09
UW CAMP 392 Hz	2.72	SOECIC MTB PDT	2.6	49	0.532	-0.07

**Significant at the Bonferroni corrected p< 0.005 level (0.05 ÷ 10)*

5.6.5 Sensitivity to musicianship

As there were only 2 self-reported musicians within the CI group (CI 1, CI 3), no statistical analysis was conducted. However, the musicians' scores can be seen in Figure 5.15, Figure 5.16 and Figure 5.17, as shaded data points. Generally, for the PMMA and the MCI, the musicians performed well, although not at ceiling. For the adaptive tests however, musicians' performance was very mixed, and they were not necessarily the best performers.

5.6.6 Summary

In summary, both the good and poor performer examples of the MedEl MuSIC Test and the SOECIC MTB PDT were shown to have a sufficient number of successful trials in order to keep $p < 0.05$. This was also true for the good performer example in the UW CAMP, however the poor performer example was shown to have an insufficient number of trials in order to keep the likelihood of achieving such a score to less than 5%.

Floor and ceiling effects were not big issues for the majority of tests, however the MCI 1 semitone interval caused floor effects for 3/15 CI users, and the UW CAMP 330 Hz caused ceiling effects in 4/15 and the 392 Hz caused ceiling effects in 3/15 CI users. No floor or ceiling effects were seen in the SOECIC MTB PDT with CI users.

The UW CAMP 262 Hz and the 5 intervals of the MCI tests were the only tests that met the criteria for reliability.

The MedEl MuSIC Test was shown to have significantly poorer scores compared to the SOECIC MTB PDT and all 3 base notes of the UW CAMP. The UW CAMP 262 Hz was shown to be significantly poorer than the 330 Hz. No statistical comparisons were made, however the 2 musicians were not shown to be the best performers across the tests.

5.7 Discussion

The overall aim of this chapter was to evaluate each of the tests to determine their suitability for use with CI users, in terms of validity, reliability and clinical use. This experiment used both NHL and CI users; the use of NHL was beneficial as it allowed the tests to be trialled with NHL prior to use with CI users. It was generally hypothesised that the tests would perform differently to each other across the predetermined assessment criteria; and that some tests would demonstrate strengths in some areas and weaknesses in others.

5.7.1 Do the tests provide enough trials to keep chance to a minimum?

The aim was to establish whether the tests provided enough trials to ensure that chance was not affecting the results: to keep the probability that the result was due to chance below 5%.

MCS

All 4 of the MCS tests provided enough trials to ensure that it was possible to score 'successfully' and to keep $p < 0.05$. The PMMA presented 40 pairs, the MACarena presented 24 pairs and the MBEA presented 31 pairs. The MCI was not so straightforward: Galvin, Fu and Nogaki, (2007) state a varied number of repeats were used until participants reached stability (p. 307), and then the average score was calculated. This was at least 2 repeats, but was typically more than 5. In the current experiment, 3 repeats were used, resulting in 27 trials per interval.

These tests presented a mixed level of difficulty: the PMMA compared 2, 3, 4 and 5 notes, with interval differences ranging from 1-12 semitones; the MBEA compared 2 musical phrases with interval differences ranging from 1-6 semitones; the MCI presented contours with interval ranges spanning 2-4 semitones (for the 1 semitone condition) and 10-20 semitones (for the 5 semitone condition); and the MACarena compared two notes with intervals between them of 1 and 2 semitones. These mixed levels of difficulty are unsurprising: each test was designed to test a range of abilities, however it makes the determination of ability regarding specific pitch intervals much harder. Information could be obtained from the raw results, however this would be time consuming, not practical in a clinical environment and the number of repeats per interval would be reduced, which would increase the likelihood of chance affecting the results.

These tests do not provide a 'cut off' score for success, although the PMMA does report 'rank norms', relating to the abilities of the normative data obtained by Gordon (1979), and the MBEA reported 78% as being 2 standard deviations below the mean (of the 'composite' score: average of all 6 MBEA tests), and this 'low score' was used as an indicator of amusia (Peretz, Champod and Hyde, 2003; Mandell, Schulze and Schlaug, 2007). The importance of enough successful trials has been discussed by Gfeller *et al.* (2002), with the requirement of achieving 9/11 correct responses, in order to achieve statistical confidence at $p < 0.05$.

Within the current experiment, four NHL (NHL 3, 15, 16, 23) scored below 78% on the MBEA scale test. Scores of 78% achieved across all components of the MBEA could signify amusia (Peretz *et al*, 2003), and whilst the scale test alone cannot be used as a diagnostic tool in this way, it could be indicative of poor performance that could signify amusia. These 4 NHL performed well on all other tests, and one of these (NHL 3) was a musician, so it may be that the MBEA was harder than the other tests. All the NHL knew that they were taking part in research that was related to CIs, and so may have not expected the subtler 'differences' that were presented in the MBEA. Compared to the presentation of pairs of 2, 3, 4 or 5 notes seen in the PMMA and MACarena, the MBEA presented two phrases of notes which averaged 5.1 seconds long, with only one 'different' note. Therefore, it is possible that this note was easily missed within a 5 second musical phrase.

The MACarena and the MBEA were not used to test the CI users within this experiment. The MACarena was felt to be too limited by the interval choices of only 1 and 2 semitones: the usefulness of these intervals would be limited for CI users, plus the number of each interval was limited and unbalanced (with 14 of the 16 'different' trials testing 2 semitones and 2 testing 1 semitone). The MBEA showed no ceiling effects and could be considered to be challenging to NHL, it was felt that it might be too challenging for CI users, indeed, previous use with CI users has shown results at chance level (Cooper, Tobey and Loizou, 2008; Cullington and Zeng, 2010). As such, of the MCS tests, only the PMMA was used to test CI users.

Adaptive tests

It was initially thought that adaptive tests in general would not provide high enough numbers of repeats easily, because they cover a wide range of intervals with minimal repeats, in order to increase efficiency. However, they are designed to maximise the number of trials around asymptote, and generally, this was seen across different types of performer for NHL and CI users. Not every participant was analysed in detail: a good performer and a poor performer were selected per test, and so the conclusions drawn here should be considered with that in mind.

For CI users, the MedEl MuSIC Test and the SOECIC MTB PDT were shown to provide enough trials to keep $p < 0.05$, for the intervals at or close to the final score. The UW CAMP also showed sufficient trials for the good performer, however the poor performer did not have sufficient trials. For the NHL, the UW CAMP showed sufficient trials for both good and poor performers, the MedEl MuSIC Test showed sufficient trials for the poor performer, however did not show sufficient trials for the good performer.

The insufficient trial numbers for the UW CAMP for the poor CI performer seemed to be due to the fact that the third run was so different to the first 2 runs (Figure 5.12 Figure 5.12), a problem that may not have been foreseen at the design stage of the UW CAMP. However this unusual 3rd run response may well be due to a non-monotonic element to CI 2's psychometric function, or may be due to a lapse in concentration or judgement. The insufficient trial numbers for the good performer on the MedEl MuSIC Test can be explained by the algorithm terminating due to the fastest possible descent of the staircase, with 3 successful trials at 22, 11 and 1 quartertone. This is a unique situation amongst these adaptive tests: the UW CAMP and the SOECIC MTB PDT maintain the same algorithm termination regardless of performance (e.g. terminating after 8 and 7 reversals, respectively). The assumption made by the MedEl MuSIC Test seems to be that if the listener can achieve those successful trials, their threshold must be 1 quartertone and further

testing is not required, however the problem here is that it is possible to achieve such a result by chance alone, and this is greater than $p = 0.05$.

The SOECIC MTB PDT provided sufficient trials for both good and poor performers for CI users, appearing to make it superior to the UW CAMP and MedEl MuSIC Test. A number of differences between the tests are likely to be in part responsible. Firstly, the SOECIC MTB PDT was the only test to use a 3AFC procedure, meaning that the chance level was 33% rather than 50%, and therefore the binomial calculator showed that fewer successful trials were needed to keep the level of chance below 5%. In addition, the number of trials per interval differed as a result of the staircase rule: the UW CAMP used a 1 down 1 up staircase, meaning that less repeats occurred per ascent or descent of the staircase, compared to the SOECIC MTB PDT (2 down 1 up) and the MedEl MuSIC Test (3 down 1 up), however the UW CAMP was the only test to repeat each staircase 3 times and then take an average, therefore increasing the trial numbers per interval that are included in the final score calculation.

Each test's algorithm terminated using different rules: UW CAMP after 8 reversals and final score the average of the last 6; SOECIC MTB PDT terminated at 7 reversals and final score was the average of the last 5; MedEl MuSIC Test terminated at 8 reversals. The calculation of the MedEl MuSIC Test final score is not certain: it appeared to approximate the average of the last 5 reversals, however even when the average of the last 5 reversals was rounded up or down, this was not an accurate calculation. Much effort has gone into calculating possibilities using the data from this work, plus repeated emails to the test creators which have unfortunately not resulted in any clear answers. Additionally, the SOECIC MTB PDT used the 'easier' task of pitch discrimination rather than ranking (Maarefvand, Marozeau and Blamey, 2013; Yitao and Li, 2013) which may have affected the progression and thus affected the trial number in a particular interval region.

Finally, the SOECIC MTB PDT presented intervals as small as 1 cent, and as this was well below the capabilities of both CI users and NHL, the inclusion of an interval that was below the capabilities of any test user meant that the algorithm could terminate at 7 reversals, unlike the UW CAMP that utilised a reversal at '0' semitones in order to follow the termination algorithm, or the MedEl MuSIC Test that terminated early with perfect performance.

A particular issue arose whilst analysing the UW CAMP data. The smallest interval the UW CAMP tests is 1 semitone, yet a perfect performer can yield a test result of 0.5 semitones. This confusion is caused by the UW CAMP's automatic reversal at '0' semitones that it adds whenever a participant answers correctly at 1 semitone, thus perfect performance (illustrated in section 5.3.3) leads to a final 6 reversals of: 0, 1, 0, 1, 0, 1, resulting in a final score of 0.5. Nimmons *et al.*, (2008) advise that this is interpreted as 1 semitone, which is in agreement with the number of successes:

5/5 correct at 1 semitone, however this has led to confusion, as seen in Maarefvand, Marozeau and Blamey, (2013), where their star performer is repeatedly reported to have achieved thresholds of 0.5 semitones for every frequency tested, and in Drennan *et al.* (2015) who also stated that the minimum possible interval in the UW CAMP was 0.5 semitones.

A further issue for the UW CAMP is that of comparison between perfect performance, and very good performance. A perfect performer is awarded a final score of 0.5 semitones, interpreted as 1 semitone, whereas a very good, but not perfect performer is awarded a final score 1 semitone, interpreted as 1 semitone: they cannot be differentiated. In the example given in Figure 7.13, the final 6 reversals totalled 6, and the final score = 1 semitone. However, if this is analysed in more detail, it is apparent that this individual was asked to pitch rank the interval 2 semitones on 3 occasions, and was successful on every occasion, whereas they were asked to pitch rank 1 semitone 6 times, and was successful on only half of those occasions, indicating that this individual's ability may actually lie somewhere between 1 and 2 semitones.

It would appear that the addition of the reversal at 0 is having a negative impact on distinguishing between good performers and perfect performers, and as a result, giving an unfair advantage to good performers. It is this author's suggestion that rather than counting the 0 as one of the reversals when calculating the final score, it should be calculated as 1, which is the best possible score on the UW CAMP. That way, perfect performance would be 6/6 = 1 semitone, and the example from Figure 7.12 would be 8/6 = 1.33 semitone, which better represents each level of success, plus is no longer misleading that the score was less than 1 semitone.

The implications of the above findings are that chance may be playing too high a role in the progression of these adaptive tests, especially with CI users. Arguably, adaptive tests are not the most suitable method for CI users due to the potential for CI users to have non-monotonic psychometric functions, however their efficiency means that it is likely that they will continue to be used with CI users, and as such, suitable numbers of trials should be ensured, to try to minimise the associated error. In addition, the miscalculation of the UW CAMP scores may be causing further error and poor discriminability between good and perfect performers, and given the far reaching usage of the UW CAMP, this is important (Nimmons *et al.*, 2008; Jung *et al.*, 2009; Kang *et al.*, 2009; Won *et al.*, 2010; Jung *et al.* 2012, Maarefvand, Marozeau and Blamey, 2013, Drennan *et al.* 2015).

5.7.2 Do the tests show suitable difficulty?

Suitable difficulty was assessed by looking at the levels of floor and ceiling effects within the tests. Ceiling and floor effects are a good way to measure whether a tool is suitable for the population it is testing in terms of difficulty. A test that only shows ceiling effects cannot give a true measure of ability, nor can it inform regarding any improvement in ability. It was hypothesized that within the CI user group, both floor and ceiling effects would be seen, and that ceiling effects only would be seen for the NHL.

Initially, > 80% was considered to be a ceiling effect for MCS tests, however it wasn't then easy to use a comparable cut off for the adaptive tests, and so a simpler approach was taken. Implications of defining ceiling as 100% meant that participants scoring 90% or 95% may also suffer from the effects of being at ceiling e.g. the test is not suitable and improvements cannot be accurately measured, meaning that some tests in this analysis may suffer from ceiling effects to a greater extent than this analysis indicates.

The difficulty of these tests was generally not suitable for the NHL group: ceiling effects were seen in every test except for the SOECIC MTB PDT. This is unsurprising as the tests used here were designed for use with children, amusics or CI users, with smallest intervals of 0.5, 1 and 2 semitones, and as such NHL ability surpassed the capabilities of these tests. Typically, ceiling effects were seen in over half of the NHL group for each test, whereas only 5 NHL had ceiling effects with the MBEA. This suggested that the NHL group found the MBEA more challenging than the other tests. The NHL expectations about the difficulty of these tests may have played a role in these poor results, as they knew that this project was related to tests for CI users, and they therefore may have expected these tests to be very easy. Also, because the MBEA presented one different note amongst 4 bars of notes, it may have been the most likely to have been affected by lapses in concentration. The SOECIC MTB PDT did not show any ceiling effects because the smallest interval tested was 1 cent, an interval that may not even be able to be heard out by the best performing musicians. No floor effects were seen for any of the tests with the NHL group.

The implications of these results are that these tests, with the exception of the SOECIC MTB PDT and the MBEA, in their current form, are not suitable for use with NHL. The ceiling effects mean that changes in ability cannot be shown accurately and the true magnitude of ability cannot be demonstrated, and so tests with a more appropriate level of difficulty should be used instead, when testing NHL. Using the same test in order to compare or to show the difference between NHL and CI users may have some merit, to demonstrate the areas of difference and similarity, however the lack of true magnitude estimation which occurs with floor and ceiling effects means this can only be a rough comparison at best.

These tests typically showed an appropriate level of difficulty for CI users, with very few floor and ceiling effects being seen. As with the NHL, no floor or ceiling effects were seen for the SOECIC MTB PDT. The MCI was considered as a whole, so looking at all 5 intervals there was only 1 person effected by ceiling effects, and 5 people affected by floor effects, however there were no floor effects at 5 semitones and no ceiling effects at 1 semitone, indicating that the MCI was not limited by these factors. The MCI covers a wide range of difficulties both within its differing contours and also its range of intervals, which means that in terms of difficulty, the test is ideally suited to testing CI users.

The UW CAMP showed the greatest number of ceiling effects, with 8 people performing at ceiling: 262 Hz had 1 CI user at ceiling, 330 Hz had 4 CI users and 392 Hz had 3 CI users. The literature indicates some CI users can succeed on pitch tasks with intervals smaller than 1 semitone. Gfeller et al's (2002) study of the pitch ranking ability of CI users didn't test intervals below 1 semitone, and reported a range of ability starting at 1 semitone, indicating a ceiling effect. Van Besouw and Grasmeder (2011) reported CI user pitch discrimination abilities of < 1 semitone. This is likely to be in part why the UW CAMP suffered from ceiling effects. In addition, the adaptive procedure may be in part responsible: the 1 down 1 up methodology may allow for a less stringent approach and may allow the influence of chance to be too great, as discussed in section 5.7.1 above. In addition, the issues with the reversal at zero, resulting in perfect performers scoring 1 semitone (UW CAMP 'reported' scores of 0.5 semitone) and also good, but not perfect performers also scoring 1 semitone (UW CAMP scores of 0.68 or 1 semitone) may mean that more participants were scoring 1 semitone than should have been; the possibility of rescoring using 1 semitone as a reversal in the averaging rather than 0 may improve this, and should be considered as a line of future work. The UW CAMP only had one CI user with a floor effect at 262 Hz, suggesting that upper interval of 12 semitones is suitably large to encompass most CI user abilities.

Whilst the MedEl MuSIC Test performed well, the floor effect seen highlighted problems with the test's methodology. CI 1 achieved an overall score of '0' (quartertones) for both T1 and T2. Looking in closer detail, (Figure 5.18), CI 1 could not achieve 3 correct responses in a row successfully enough to progress down the staircase, and as a result, the interval kept increasing, which continued until the presented pair of notes were 262 Hz with C8 (96 quartertones, or 48 semitones). This did not happen for the other 14 CI users in the study, so it doesn't appear to be a typical downfall of the test in general, and may reflect some phenomenon particular to CI 1. CI 1 performed suitably well across the other tests, so rather than being a clear deficit in his ability, it may relate to internal criterion values, perhaps at the third repeat of an interval pair he started doubting himself and chose the opposite answer. It is also possible that some CI users may have

mistaken a repeated pair as an indication that their first answer was wrong, and changed their answer accordingly, although the ability to be sure that the second (or third) pair was identical to what they had just heard may be a fairly difficult task. This is possible with people that may have little prior knowledge of research or psychoacoustic methods, which is unlikely to be the case for the NHL group for example, who were primarily recruited from the university environment.

This example also highlights the differences that would have been seen in the result if, for example, the UW CAMP's requirement of only needing 1 correct answer was utilised instead of a 3 down 1 up procedure: the staircase would have started to descend and the participant may well have achieved a much 'better' score as a result: indeed CI 1's scores for the UW CAMP ranged from 1.28 – 1.78 semitones. The MedEl MuSIC Test showed no ceiling effects, which could be interpreted as 0.5 semitones being a small enough interval and a good level of difficulty, but this may well be affected by the requirement of 3 correct responses in a row that are required to descend the staircase.

With such wide pitch intervals in the MedEl MuSIC Test, loudness may have also played a part either in facilitating or hindering the responses, and it was noted by the experimenter at the time that the differences between those notes resulted in subjective loudness differences. For the majority of CI users and NHL, the intervals were not as large and so differences in loudness were considered a much smaller issue. The papers that report the MedEl MuSIC Test do not mention loudness issues, and it may warrant further investigation given the potential widespread use of the MedEl MuSIC Test.

There is no reporting of any scores of 0 in the MedEl MuSIC Test literature, however, Brockmeier *et al.*, (2011) report that the range of abilities seen in their CI user group was from 1 quartertone upwards for all timbres, indicating that their participants were performing slightly better than the CI users within the current Study 3. It is not clear from their paper how many participants scored 1 quartertone (0.5 semitone), however as the medians were 20.6, 16.7 and 11.5, it is unlikely to be many. The results from Brockmeier *et al.*, (2011) do indicate that 0.5 may not be a small enough interval for testing CI users.

Implications of high levels of ceiling effects (and potentially miss-scoring issues) are that the UW CAMP in its current state is not suitably difficult to measure the true ability of CI users, and will be limited in its ability to measure change. This is particularly concerning given the wide and continued usage of the UW CAMP (Nimmons *et al.*, 2008; Jung *et al.*, 2009; Kang *et al.*, 2009; Won *et al.*, 2010; Maarefvand, Marozeau and Blamey, 2013).

The results of the current study would indicate that 12 semitones is a large enough interval to include as an upper value, and that the lower value should be below 1 semitone, but also that the importance of the intermediate intervals should not be disregarded. In addition to the interval sizes, the difficulty of a test can also be affected by the speed and complexity of the pitch materials, as well as the methodology used to test.

5.7.3 Do the tests show suitable reliability?

Reliability was reported in the current study using the ICC, and suitable reliability was defined by a 'good to excellent' ICC of greater than or equal to 0.8 (Pinna *et al.*, 2007), and in addition an ICC value that was significantly different from what was expected by chance, using the Pearson's critical value of *r* for *n*. Arguably the cut offs for these benchmarks in reliability strength are arbitrary, however the higher the coefficient, the more of the variability can be attributed to genuine differences and less due to test error. It was hypothesized that at least some of the reliability coefficients would not be high enough to meet the criteria of ≥ 0.8.

Of the tests used with CI users, only the MCI and the UW CAMP Hz and 330 Hz tests achieved an ICC of ≥ 0.8. The best performing tests were the MCI 4 and MCI 3, with ICCs of > 0.9. It is thought that the relative ease of the task with intervals of 3 and 4 semitones may have influenced these results, plus the MCS meant that there were no adaptive decisions that may introduce error. It would therefore be expected that the 5 semitone interval might perform even better for reliability, which wasn't the case. In this experiment, the MCI was always started with 5 semitones and this may have been the introductory phase as the CI users adjusted to the task. When the intervals were reduced to 2 and 1 semitone, the reliabilities dropped, which might be explained by the task becoming more difficult, however this was not reflected in the median scores (Table 5.7).

If the reliability criteria was relaxed to ≥0.75 rather than ≥0.8, all tests except for the SOECIC MTB PDT would be considered reliable enough; indeed ICCs of greater than 0.75 are often considered to be acceptable and are argued to be 'good' (Fitzpatrick *et al.*, 1998; Koo and Li, 2016). This means that 75% of the variability (or greater) can be attributed to genuine differences, and 25% (or less) to error within the test methodology. There was no clear distinction between MCS and adaptive tests in terms of reliability coefficients in this study.

In addition, for the UW CAMP and the MedEl MuSIC Test, test users must be able to pitch rank in order to succeed, rather than discriminate, making the test more difficult (Maarefvand, Marozeau and Blamey, 2013; Yitao and Li, 2013). In addition to this, the way in which the adaptive tests progressed through their staircases and had their scores calculated was different. It is also possible

that some of the CI users in this study had non-monotonic psychometric functions and therefore were not progressing along the adaptive staircase as expected, which would influence the results.

Comparisons to the literature are not easily made, as although the MedEl MuSIC Test was assessed for test retest reliability (Brockmeier *et al.*, 2011), it was reported that no significant difference was seen, when 9 CI users were tested 6-12 weeks later, however no further details were given. Kang *et al.*, (2009) tested the reliability of the UW CAMP using the ICC and reported an (undefined) ICC of 0.85 for the UW CAMP, which may be relating to all the base notes combined, as further details were not specified. This is in keeping with the ICCs reported here.

The poorest performer was the SOECIC MTB PDT, with an ICC value of 0.43, and a significance test showed that this could not be distinguished from the critical value of *r,* which represented the correlation coefficient that might be expected to occur by chance. This has quite substantial implications for the use and validity of the SOECIC MTB PDT with CI users. Whilst the SOECIC MTB PDT is a pitch discrimination task, and is considered easier than the (pitch ranking) MedEl MuSIC Test or UW CAMP tests, unlike those tests, which reduce their step size to 1 quartertone and 1 semitone respectively, the SOECIC MTB PDT continues with the predetermined step sizes. This means that if one of the 5 reversals used to calculate the final score happens to be in an area of large step size (e.g. 4, or 8 semitones) and the other reversals are much nearer to 0.64 or 0.32 semitones (which are the first two values below 1 semitone chosen by the SOECIC MTB PTD's designers), the larger reversal value is going to have a huge impact on the final result.

In addition, the SOECIC MTB PDT presented intervals that could be either above or below the target note: e.g. with a base note of F4 and an interval of 4 semitones, the interval presented for the two trials could be formed by F4 and A4 (e.g. F4 + 4 semitones), and then F4 and C#4 (e.g. F4 – 4 semitones). For CI users, and likely for some NHL, these two intervals may be considerably different in their difficulty. This means that the adaptive procedure requiring two correct responses in a row might behave more like a procedure that requires 1 correct response, but for some intervals only, whilst scoring as though it is using a 2 down 1 up procedure. This also means that there are less repeats per unique interval, meaning that the likelihood of chance affecting the results for that interval is increased.

NHL performed at ceiling for the majority of the pitch tests and so conducting the test a second time would not be useful and test retest reliability could not be accurately calculated. As the SOECIC MTB PDT test was unaffected by ceiling effects, it was used in Study 2 to assess reliability on retest. Initial results indicated a strong reliability, however the graphical display of the data indicated that 2 NHL points may have been biasing the result. It was suspected that these 2 individuals may have some form of amusia, it is estimated to be present in 4% of the population

Peretz, Cummings and Dubé (2007), and although each individual was screened for this, they may have been unaware of their level of ability, if it has never been formally tested. In hindsight, the questions regarding self-perception of tone deafness prior to recruitment may not have been sufficient.

Removal of the two outliers resulted in a much reduced ICC, no longer considered to be strong. It is thought that the possible reason for this is due to the fixed step sizes, which means that depending where the final 5 reversals fall, the average score may be much more affected if the final reversals cover a large step size area, compared to a smaller step size area. It was thought that the fixed step sizes may have been responsible for the poor ICC with NHL, which was also seen with the CI users. Previous work has shown correlation coefficients of 0.69 between T1 and T2 with 18 NHL (Lamb, 2011), and did not discuss these issues.

5.7.4 Do the tests demonstrate concurrent validity?

To establish validity of the tests, the results achieved in the current study were compared to the literature. Similar findings to the literature would establish that the tests were administered in a similarly (appropriate) way and would indicate that the populations of CI users and NHL were similarly sampled. Initially this research question was included to investigate whether the tests demonstrated concurrent validity, however the vast differences in approach and methodology meant that it was hypothesized instead that significant differences would be expected between the tests, rather than similarities. This analysis was felt to be important because clinicians may be inspired to compare recent pitch perception test results with historical ones, even if the tests were different, without full appreciation of just how different the methodologies might be. To date, no previous work has compared the results of different pitch perception tests using the same CI users.

The scores achieved by the current group of CI users were in keeping with previous literature (Gfeller and Lansing, 1992; Galvin, Fu and Nogaki, 2007; Lassaletta *et al.*, 2008; Nimmons *et al.*, 2008; Jung *et al.*, 2009; Kang *et al.*, 2009; Brockmeier *et al.*, 2011b; van Besouw and Grasmeder, 2011). The current MedEl MuSIC Test results were slightly higher than previous published work, this may have been due to individual differences in participant groups, as Brockmeier *et al.*, (2011) used 31 MedEl MuSIC Test CI users prior to 2011. In addition, their application of the test was slightly different to Study 3, as they used a pure tone, a piano tone and a string tone interweaved. The application of the MCI within Study 3 was also different to what was presented in Galvin, Fu and Nogaki, (2007) as the current study used 3 repeats and blocked the trials by interval size whereas the Galvin *et al.* study repeated the MCI until the results stabilised, and interweaved the

intervals. The scores achieved by the NHL group were also in keeping with the literature (Schuppert *et al.*, 2000; Peretz, Champod and Hyde, 2003; Jung *et al.*, 2009; Kang *et al.*, 2009; Brockmeier *et al.*, 2011). Although it was expected that differences would be seen across the tests, some weak evidence was discovered indicating that actually, some NHL and CI users do demonstrate similar good and bad results across a number of the tests used. CI 9 performed outstandingly across every test, with scores of around 90% for the PMMA and MCI tests, and threshold scores of 2, 1 and 0.62 semitones. This provides further evidence that 1 semitone is not a difficult enough interval for every CI user. The UW CAMP cannot test smaller than 1 semitone, and so the ceiling effects seen here indicate that the UW CAMP is not informative for this CI user, and the poor reliability issues seen with the SOECIC MTB PDT sheds doubt upon the SOECIC MTB PDT result, even though the range of difficulty is superior with the SOECIC MTB PDT.

CI 15 showed similar scores to the CI median scores from this study for the UW CAMP, MedEl MuSIC Test and SOECIC MTB PDT, however showed much poorer scores for all parts of the MCI. Throughout this study, tests that use MCS (e.g. the PMMA and MCI) have generally shown improved and more reliable results than adaptive measures, however CI 15 is an unusual example in that he performed much better with the adaptive tests rather than one of the MCS tests. A possible reason for this might be the speed at which the MCI notes were presented, each note was 250 ms with an interval of 50 ms, meaning that presentation of 5 notes happens within the space of 1.45 seconds. This, coupled with the smallest intervals (1 and 2 semitones) could be the cause of CI 15's poor results, which for 1 and 2 semitones are below chance level.

In addition to the two CI users described above, two NHL also showed interesting results across the tests. The two NHL were the poorest performers with the MedEl MuSIC Test, and showed poorer scores across every test, compared to the NHL medians for each test. Despite the huge differences in test approach and methodology, it was interesting to see that these poorly performing NHL did badly across the majority of tests: a weak form of concurrent validity. Both did very poorly on the MedEl MuSIC Test, achieving thresholds of 9 and 11 semitones, and generally did much worse than the average score for the adaptive tests, except for the UW CAMP 330 Hz, which they performed at a similar (good) level to the average score.

Both these NHL were non-musicians, and it may be that they have aspects of amusia. The MBEA demonstrates evidence of amusia if the composite score (average of all 6 MBEA subtests) is less than 78% (Peretz, Champod and Hyde, 2003). Neither scored less than this for the scale subtest, however if they were to complete the MBEA entirely, they may meet this criteria. This may indicate that NHLs with amusic-style problems might also theoretically have difficulty with adaptive procedures in a similar way to CI users. Perhaps, like CI users, it cannot be assumed that people

with amusia have monotonic psychometric functions for pitch intervals, although this was not reported by Foxton *et al.* (2004) who estimated the psychometric curves in 10 self-reported amusics.

Finally, comparisons were made between the UW CAMP, MedEl MuSIC Test and SOECIC MTB PDT CI user scores, as they all measured the pitch 'threshold' in semitones. It should be noted that whilst the UW CAMP and MedEl MuSIC Test measured pitch ranking threshold, the SOECIC MTB PDT measured the pitch discrimination threshold. Results showed significantly poorer scores with the MedEl MuSIC Test, when compared to all UW CAMP tests and the SOECIC MTB PDT.

A main difference between these 3 tests was the differences in the probabilities of a positive response at convergence (Levitt, 1979): the MedEl MuSIC Test requires 3 correct responses in a row and therefore the probability of a positive response at convergence is 79%, the SOECIC MTB PDT requires 2 correct responses and the probability is 71%, and the UW CAMP requires only 1 correct response and the probability is 50%. The accuracy of this has been questioned in NHL (García-Pérez, 2014), and so in CI users where the use of adaptive methods has been questioned (Swanson, 2008; Maarefvand, Marozeau and Blamey, 2013) this may be even less accurate. If the curve was very steep, it might be expected that significant differences would be seen between tests using different tracking locations; however it might be expected that a test that tracks at 50% would show a poorer score than one that tracks at 79%, and this is the reverse of what was shown in Study 3. These comparisons are also clouded by the differences in base notes across the tests.

The MedEl MuSIC Test also had the widest range of note intervals and as such also suffered the most from loudness differences (as measured with a hand held SLM) and in addition, it was unclear how the final result was calculated. These factors may also influence such a poorer average performance, in addition to the possible self-doubting bias that might affect CI users when asked to repeat the same task 3 times in a row prior to the test progressing. The implications of this are that requiring CI users to achieve 3 correct responses in a row may be too difficult, and a less difficult task should be employed, if adaptive procedures cannot be avoided.

In addition, the base note 262 Hz was much worse than the 330 Hz or 392 Hz notes in the UW CAMP. This is of great interest because each test uses identical methodology, tracking and scoring, and yet, substantially different scores were obtained. The poorer result at the lowest tone is not a pattern that has been reported in the literature with CI users; Kang *et al.*, (2009) reported similar scores for all 3 base notes (between 2.5 – 3.4 semitones) whereas Jung *et al.*, (2009) reported a much poorer average for the highest note, 392 Hz (8.1 semitones). Nimmons *et al.*, (2008) didn't report averages but showed that 262 Hz had the widest ranging scores (1-11.5 semitones).

It is possible that the effects of different frequencies may be responsible for these results, however between Study 3 and Nimmons *et al.*, (2008); Jung *et al.*, (2009); Kang *et al.*, (2009) there were great differences in which frequencies were the best and worst performers. The base notes also are very close to each other, thus a threshold of greater than 3-4 semitones overlaps the next base note up, and so these notes are not spread out enough to be independently informative about different areas of the cochlea. Alternatively, these results may reflect test error or random noise, or maybe they do reflect differences due to frequency within the implanted cochlea, but that individual differences mean averaging the data across many CI users is not an appropriate method of analysis.

The implications of this are that tests cannot and should not be used interchangeably, just because they are all tests of pitch: when any test is used in a clinical setting, every variable should be clearly noted, as even identical tests that use a different base note can cause hugely variable results. The current study's results with the MedEl MuSIC Test provide an example of this. It is not clear whether the MedEl MuSIC Test is performing at a poorer level than the other tests; it may be that it is more accurate than the other tests, however the evidence favours the other tests as there are more of them in agreement. Therefore, it is important that any scores obtained with the MedEl MuSIC Test, or read about in the literature are considered within the boundaries of the test's limited ability to agree with other similar tests.

5.7.5 Do the tests show significant differences between musicians and non-musicians?

Known differences within a spectrum can provide validity, if those known differences can be demonstrated by using a test. For example, a valid pitch test should be able to differentiate between those very good at pitch tasks and those that are not, and an example might be musicians and non-musicians. All participants were asked whether they considered themselves to be musicians, whether they had any formal music training, and whether they participated in any regular musical activities. Musicianship was defined by answering yes to either of the first two questions, as only including formal qualifications would fail to identify a number of musically able people, and verbal self-report is a common method for this assessment (Law and Zentner, 2012). It was hypothesized that the musicians would perform significantly better than non-musicians across the tests, again providing some evidence that the tests were demonstrating criterion validity, and construct validity in the form of discriminant validity.

Within the group of 15 CI users, only 2 considered themselves to be musicians, meaning that no statistical comparisons could be carried out. Within the 23 NHL, 11 considered themselves to be musicians and/or had formal music training. Even though ceiling effects were present within the

group, statistically significant differences were seen between the musician and the non-musician group for every test except the MBEA (see Table 5.6). Ceiling effects occurred for the musicians rather than the non-musicians, which explains why the differences were still seen.

The MBEA did not show differences between the groups, and this may be because the musicians did not score as well on the MBEA as they did on the other tests. Generally, high scores and ceiling effects are not unseen with the MBEA: Peretz, Champod and Hyde, (2003) showed around 65 of the 160 NHL tested scored average of above 90%. It is possible that the NHL within the current study found the tests other than the MBEA in this study much simpler. The MBEA presented a 4-bar phrase lasting 5-7 seconds long, with only one note different between the phrase pairs, and in addition, the notes were presented at a fast pace. Alternatively, it may be that the MBEA was less sensitive to differences in musicianship, and this is not something that has been looked at previously with the MBEA and CI users, although it was mentioned as a source of variability by Cooper, Tobey and Loizou, (2008). The Montreal Battery for the Evaluation of Musical Abilities (MBEMA), which was a further development from the MBEA in order to test for amusia in children under the age of 10, has shown that musical training does significantly affect scores (Peretz *et al.*, 2013).

There are no large studies comparing CI user musicians and non-musicians: it seems likely that this is because CI user populations do not tend to have a huge number of musicians within them. Gfeller *et al.* (2008) conducted a multi-faceted assessment of various approaches to music assessment for CI users and showed that high school or adult (but not younger school years) music training showed correlations with tests of pitch ranking, familiar melody recognition, and timbre recognition. In addition, Chen *et al.*, (2010) provided evidence that the longer children with CIs have had musical training correlated with pitch discrimination and ranking performance with real world piano tones. Maarefvand, Marozeau and Blamey, (2013) reported the musical abilities of a star performer, who had musical experience pre and post implantation, who achieved ceiling effects using the UW CAMP pitch test at the original frequencies plus at extended frequencies higher and lower.

These results suggest that the PMMA, MedEl MuSIC Test, MACarena, SOECIC MTB PDT and UW CAMP are all sensitive to known differences, in the case of musicians vs. non-musicians within a group of NHL. This supports criterion validity within these tests, for NHL. It was not possible, within Study 3 to assess these differences within the CI user group, due to the low number of musicians, however existing literature using a correlational approach has indicated that CI users with greater musical experience often show better results on pitch based tests.

5.7.6 Summary: suitability of the tests for CI users

In terms of numbers of trials, the MCS tests performed well, with enough trials repeated to have confidence in the result. It would be helpful to have further guidance regarding:

1. a definition of success e.g. a score of 75% (for example) is required to state that the interval of 3 semitones has been successfully pitch discriminated or pitch ranked
2. the difficulty assessed by the test.

The adaptive tests did not always provide enough trials or accept enough trials when calculating final scores, to keep the alpha value below 5%. This tended to occur for best performers: it seems that the test designers made the assumption that a good run indicating perfect performance (e.g. every trial correct) could only be made by a star performer and that chance could not have any affect, which is not the case. Final scores were often calculated based on too few successful trials; this may be solved by altering algorithms to terminate after more reversals and calculating the scores using more reversals, which is often seen in more traditional psychoacoustic methods.

The best performing test in terms of trial numbers was the SOECIC MTB PDT, likely because of the 3AFC, the chance level was 33% rather than 50%, thus reducing the effect of chance; the 2 down 1 up method maximised trials, plus using an average of 6 of 8 reversals to calculate the final score and finally the inclusion of such a wide range of intervals meant that ceiling or floor effects weren't seen and therefore didn't hinder the number of trials or the final score.

The level of difficulty of the tests was much more suitable for CI users compared to NHL. The best performing tests were the SOECIC MTB PDT and the MCI. The SOECIC MTB PDT included a wide enough range of intervals that was able to encompass the typical ability of both CI users and NHL within the same test. In addition, the MCI showed no ceiling effects at the smallest interval and no floor effects at the largest, indicating its suitability as a holistic test. Conversely, the UW CAMP showed ceiling effects for 8/15 CI users, and this had also been seen in previous work (Nimmons *et al.*, 2008; Jung *et al.*, 2009; Kang *et al.*, 2009), indicating that it does not provide a suitable level of difficulty to properly assess the baseline of CI users' ability, nor any improvement.

The best performer was again the SOECIC MTB PDT with a very wide ranging number of intervals, from 1 cent (0.01 semitone) to 16 semitones. This allowed plenty of scope for testing CI users below 1 semitone, and also was able to test NHL including star performing musicians. The MCI also performed very well, showing no floor and ceiling effects at 5 and 1 semitones, respectively.

Reliability level was mixed, with only the MCI and 262 Hz and 330 Hz UW CAMP meeting the criteria of ICC > 0.8 and being significantly different from what is expected due to chance. All tests

except the SOECIC MTB PDT achieved an ICC of 0.75, so generally performance could be considered acceptable. The SOECIC MTB PDT scored very poorly on reliability, an ICC of 0.43 was seen with CI users and an ICC of 0.04 was seen with NHL. It was thought that the fixed step sizes that are used and the possibility that the intervals are sometimes above the base note and sometimes below, may be responsible for these poor results. The best performers, with ICCs of > 0.9 were the MCI with 4 and with 3 semitones. This may be due to the redundancy seen within the MCI due to the multiple notes used to form the contour. In addition, the larger sized intervals were the easiest of the contours, disregarding the first (5 semitones) which may have been used as an acclimatisation period for the test.

Some limited evidence was seen in both CI users and NHL to indicate that overall the tests demonstrated concurrent validity: people that perform noticeably well, or noticeably poorly on one test tend to do so for all tests. Within the CI results, the MedEl MuSIC Test was shown to be significantly poorer than the UW CAMP or SOECIC MTB PDT scores, which was thought to be due to the large intervals and subsequent loudness issues, coupled with the 3 down 1 up method which is arguably too hard for some CI users. In addition, the base note 262 Hz within the UW CAMP was significantly poorer than the 330 Hz and 392 Hz. This has not been previously shown in the literature, although wide variations between the base notes have been seen (Jung *et al.*, 2009). It is possible that this reflects the effects of frequency, although rather than being a finding that can be applied to CI users in general, it may reflect individual differences in CI users. In addition, test error and individual error may play a role.

Every test except for the MBEA showed significant differences between musicians and non-musicians within the NHL. This provides some evidence of criterion validity: that a test shows 'significant relationships with external musical proficiency indicators' (Law & Zentner, 2012, pp1).

Finally, there were some theoretical adaptive issues regarding the suitability of adaptive staircase procedures with CI users, due the inappropriate assumption that CI users pitch perception abilities would follow a monotonic psychometric curve (Levitt, 1971; McDermott, 2004; Swanson, 2008; Maarefvand, Marozeau and Blamey, 2013). It is not possible to assess from Experiment 1 whether this phenomenon affected the results; and this is something that will be investigated in Experiment 2.

A tabulated summary of the ideal features is presented in Table 5.11 below (please note the reversed phrasing of the questions regarding floor and ceiling effects in order to ensure ticks represent a positive feature).

Table 5.11 Summary of test features

test		Suitable repeats?	No ceiling effects?		No floor effects?		Small enough interval for CI?	Test retest reliability		Difference between musicians and non-musicians? (NHL only tested)	Suitable (non-adaptive) method for CI?
			NHL	CI	NHL	CI		NHL	CI		
MACarena		✓	×	n/a	✓	n/a	×	n/a	n/a	✓	✓
MBEA		✓	×	n/a	✓	n/a	×	n/a	n/a	×	✓
PMMA		✓	×	✓	✓	×	×	n/a	×	✓	✓
MedEl MuSIC Test		✓	×	✓	✓	×	✓	n/a	×	✓	×
UW CAMP	262 Hz	✓	×	×	✓	×	×	n/a	✓	✓	×
	330 Hz	✓	×	×	✓	✓	×	n/a	✓	✓	×
	392 Hz	✓	×	×	✓	✓	×	n/a	×	✓	×
SOECIC MTB PDT		✓	✓	✓	✓	✓	✓	×	×	✓	×
MCI	5	✓	n/a	✓	n/a	✓	✓	n/a	✓	n/a	✓
MCI	4	✓	n/a	✓	n/a	×	✓	n/a	✓	n/a	✓
MCI	3	✓	n/a	✓	n/a	✓	✓	n/a	✓	n/a	✓
MCI	2	✓	n/a	×	n/a	×	✓	n/a	✓	n/a	✓
MCI	1	✓	n/a	✓	n/a	×	✓	n/a	✓	n/a	✓

The conclusion from Experiment 1 was that there was no clearly superior test, each had strengths and weaknesses, though the MCI was a strong contender both in design and performance. However, the MCI did not provide information specific to intervals, plus it was complex and fairly fast.

One of the main limitations to this experiment was the issue of comparing tests that are different by design. The SOECIC MTB PDT used a much wider range of base notes than the other tests, and in Study 1 all the possible notes were used, and these were chosen at random by the test. This was specified to be F4 for Studies 2 and 3, making them more comparable to the other tests used in Studies 2 and 3, but less comparable to results from Study 1.

Even with base notes specified and similar, the SOECIC MTB PDT, UW CAMP, MedEl MuSIC Test and MCI did not test the same intervals, although intervals may have overlapped across the tests, however especially with CI users, these differences in notes are likely to have had large effects on

the results, making comparisons difficult. The MCI was not included within Experiment 1 until Study 1 and 2 (with NHL) had been completed, and so there is no data for the MCI with NHL.

Chapter 6 Development of a new test: the Pitch Contour Test

6.1 Aim

To design a test:

- with enough repeats to provide statistical confidence in results
- with a suitable level of difficulty for CI users, aiming to avoid floor and ceiling effects
- with suitable methodology for CI users e.g. avoiding adaptive procedures
- that shows good reliability
- that shows differences between musicians and non-musicians
- with results that are easily interpretable for both CI users and clinicians

6.2 Features to be retained from existing tests

Features to keep:

- Ability to test pitch discrimination and pitch ranking
- MCS
- Use of the 3 AFC rather than 2AFC: this allows chance to be at 33% rather than 50%, avoids a 'same/different' task or a forced pitch ranking only task
- Different stimuli types (e.g. timbre and frequency)
- Clarity about what is being tested and how (e.g. transparency needed about interval sizes and number of repeats)

6.3 Concept design I

6.3.1 Pitch discrimination and ranking as two separate tests

The initial design included two tests, the first one was a discrimination test (e.g. an 'oddball' procedure) and the second was a pitch identification task (e.g. a ranking task), both which used the MCS. The first test used a 3 AFC procedure,

X X Y

and the participant must decide which is the odd one out. Once that test was completed, a second test, to assess pitch ranking was then presented. This test used a 2 AFC, in order to determine which note was 'higher'

X Y

Output from the test would be fed back to the participant using visual aids (e.g. a picture of a piano keyboard) to explain ability across interval sizes, stimuli types and frequencies.

Assuming the notes are audible, participants can be divided into one of the following:

Person 1 – cannot discriminate between two notes of a given interval size (discrimination and ranking = chance level).

Person 2 – can discriminate between the notes but cannot pitch rank (discrimination = good, ranking = chance level).

Person 3 – can discriminate but show pitch reversals (discrimination = good, ranking = below chance/at zero).

Person 4 – can discriminate and can pitch rank (discrimination and ranking good).

Benefits of having the discrimination test first were that:

1. The task is the easiest of the two, and so should help the participant understand the task and feel encouraged
2. If the participant cannot pitch rank, then this is a better place to start to gather basic information
3. It will identify person 1 (who shouldn't be then tested with the ranking test)
4. It is appropriate for person 1, 2 and 3.

Benefits of having the ranking test first were that:

1. If only one test is completed due to time constraints, then assuming participant is 'person 4' then ranking provides the most helpful information

2. It is appropriate for person 4 (only). Whilst the ranking test may help identify persons 2 and 3, they may be so disheartened by doing the ranking test first that they doubt their ability or withdraw consent for further testing.

Benefits of testing both pitch discrimination and pitch ranking:

1. Provides more information than either, alone.
2. Identifies between person 1, 2, 3 and 4.

In summary, the initial concept was 2 tests:

1. 3 AFC discrimination test (similar to SOECIC MTB PDT), using MCS
2. 2 AFC ranking test (similar to UW CAMP and MedEl), using MCS

6.3.2 The 'different' note positioning

The options surrounding the 3 AFC were whether to have the different note in positions 1, 2 or 3 or a combination of those. Previous work with the MCI in Chapter 5 had shown that the different complexities meant that the 9 contours had different difficulty. Parncutt and Cohen (1995) describe a useful way of expressing contour complexity, with a rising or falling contour having a complexity of 0, and a contour that changes direction, e.g. rising-falling as having a complexity of 1. This description would have to be adapted for use when describing the MCI, which also includes a flat 'contour' and rising-flat contours.

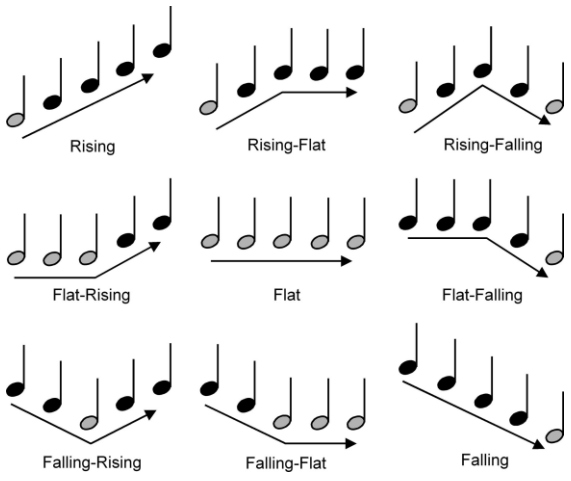


Figure 6.1 The Melodic Contour Identification test contours, from Galvin, Fu and Nogaki (2007), reproduced with permission

Because of the different levels of complexity and difficulty in the MCI, true chance of 1/9 e.g. 11% would only occur if the participant was responding without being able to differentiate between the 9 contours *at all*. If a participant could hear that a contour had a change in direction (e.g. they knew that it was not flat, rising or falling) then chance level is altered to 1/6.

To reduce this possibility, the 'different' note would only take positions 1 or 3:

O O O O

O O

And not

O

O O

If the notes had the possibility of both going higher and lower than the 2 same notes (the base notes), then there would be 4 possible combinations of pitch contour for the '3 note' discrimination test:

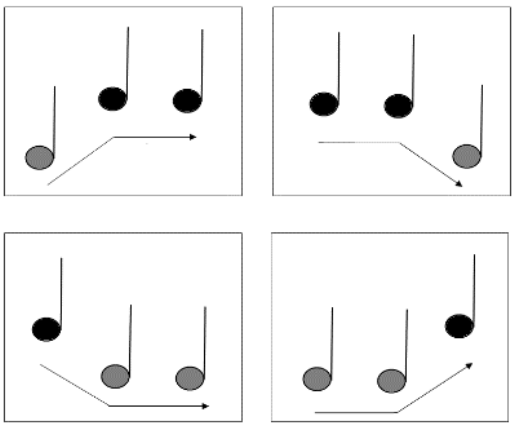


Figure 6.2 PCT contours

6.3.3 '3 note' phrase

The three note phrase is similar to the 2nd, 4th, 6th and 8th contours of the MCI in terms of rising-flat, flat-falling, falling-flat and flat-rising (see Figure 6.1). Once the flat, rising, falling, rise-fall and fall-

rise contours had been removed, the four contours that remained had the same level of complexity and potentially the same level of 'difficulty'. The 3 note phrase rather than 5 note also means that assessing the interval size is simpler and easier with the PCT rather than the MCI. In the PCT, if the interval size between notes is 1 semitone, and the base note (indicated by filled in grey, in Figure 6.2) is 'H' (the letters 'H, I, J, K and L' have been used here to represent each note, without using letters that are traditionally used to represent actual musical notes), then any of the 4 PCT contours will be testing 1 semitone, between the notes of 'H' and 'I' (e.g. H I I; I I H; I H H; H H I). If the interval size between notes in the MCI is 1 semitone, and the base note (indicated by filled in grey, in Figure 6.1) is 'H', then contour 5 (flat) will not test any interval, contour 4 (flat-rising '_/') presents notes H H H I J and therefore spans 2 semitones, and contour 1 (rising '/') presents notes H I J K L and therefore spans 4 semitones.

6.4 Concept design II

It was at this stage that the combination of information for both pitch discrimination and pitch ranking was realised: if a participant could successfully rank the pitches then they could chose the correct contour, however for participants that could only hear differences and not pitch rank, then two answers would be correct. If you hear X X Y, and you know that it is the third note that is different, then choosing either of the 2nd or 4th contours in Figure 6.2 above, would be considered the right answer for the pitch discrimination test. This meant that pitch ranking could be scored at the same time as pitch discrimination – meaning the two tests were running in parallel without any extra effort from the participant. It did mean that the chance levels were different for each test though, pitch ranking chance = 25% and pitch discrimination chance = 50%.

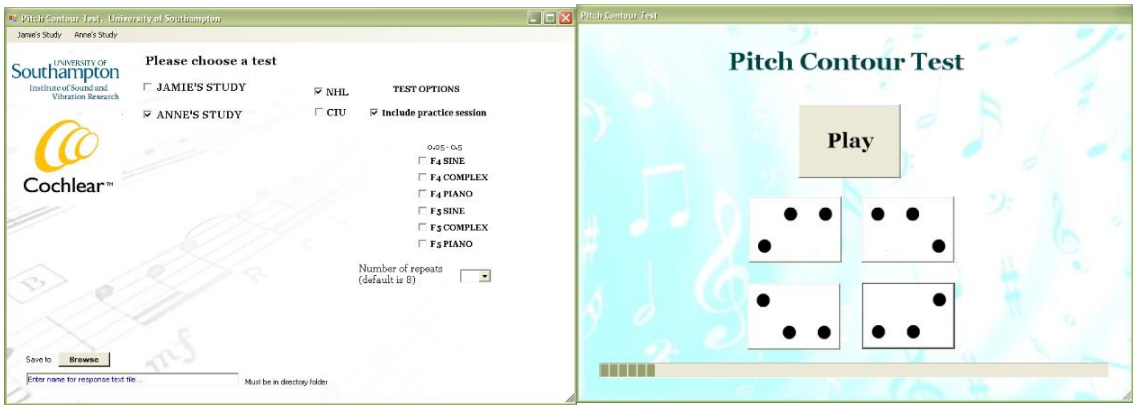


Figure 6.3 Screenshots from the PCT

First picture: front page showing test options. Second picture: the participant's interface showing the play and contour response buttons.

6.5 Design feature justification

6.5.1 Method of constant stimuli

Adaptive methods should not be used to assess the pitch perception of CI users because their psychometric curve cannot be assumed to be monotonic (McDermott, 2004; Swanson, 2008; Maarefvand, Marozeau and Blamey, 2013) which is one of the assumptions made by the up down method (Levitt, 1971). Using the MCS also allows the psychometric function to be plotted and estimated, allowing an estimation of the number of CI users affected by non-monotonic psychometric functions.

6.5.2 Triplet of notes and 4 alternative forced choice

If a test presents two notes, the participant can either be asked whether the notes are the same or different (as in the MACarena), or ask which note is higher (as in the MedEl MuSIC Test and the UW CAMP). The first option means that a larger number of trials would be needed per interval, as some of the trials also need to be 'the same'. The MedEl MuSIC Test and the UW CAMP test pitch ranking only and their design means that pitch discrimination cannot be tested in this way. The SOECIC MTB PDT presents three notes, which allows it to test pitch discrimination, and so this was the model which the PCT was based upon.

The development of the 4 contours (see Figure 6.2 above) was very similar to contours 2, 4, 6 and 8 of the MCI (see Figure 6.1). The way in which the MCI contours were presented is good: participants listen to a phrase and must chose the contour that they heard, without needing to rely on concept words e.g. higher or lower. However, having the choice of 9 contours meant that complexity and difficulty differed across the contours. The flat contour presents similarly to the 'same' trials described above, it is not gathering any information about pitch as no change in pitch is presented. Having four choices also enabled pitch discrimination and ranking to be scored simultaneously. Unlike the MCI, the PCT uses 3 notes rather than 5. This clarifies the interval size and means that each of the 4 PCT contours tests the same interval span.

6.5.3 Scoring pitch discrimination and ranking simultaneously

This is the PCT's unique selling point. The four contours make it possible to score these simultaneously, by allowing the discrimination test to count two possible answers as correct, if the participant got the correct 'odd note'. This saves time and means that more trials and stimuli can be tested. This also affects the chance level for each of the tests: pitch discrimination has a chance level of 50% and pitch ranking part has a chance level of 25%.

6.5.4 Number of trials for success

The aim was to ensure statistical confidence by presenting enough trials, to ensure that enough successful trials were used to calculate final scores. It was also considered important to allow participants to have a lapse in concentration. Due to nature of the discrimination task, which has a probability of 50%, a binomial calculator (www.stattrek.com) was used to calculate the number of trials needed to keep the probability of success below 5% e.g. an alpha value of 0.05. This was chosen to be in keeping with the typically used alpha value in statistical analysis. Examples of this probability and cumulative probability distribution can be illustrated using a coin toss repeated 3 times, assuming a head is a success (taken from www.stattrek.com).

Table 6.1 Probability and cumulative probability using a coin toss example, chance of 50% and 3 repeats

Outcome	binomial probability	Cumulative outcome	Cumulative probability of that score or greater
0/3	0.125	≥0/3	>0.999
1/3	0.375	≥1/3	0.875
2/3	0.375	≥2/3	0.5
3/3	0.125	3/3	0.125

This demonstrates that achieving 3/3 does not mean the probability of that happening by chance was less than 0.05. Initially, the idea was to provide a certain number of repeats per interval size, and so the binomial calculator was used to calculate probabilities for different numbers of successes on different numbers of trials. Four, 5, 6, 7 and 8 trials were considered, and the probabilities are calculated below.

Table 6.2 Number of trials and associated probability required to ensure chance <0.05 and allowing for one lapse in concentration.

Number of trials	Outcome	probability	Is chance less than 5%?	Does this allow for one lapse in concentration
Four	4/4	0.06	✗	✗
Five	5/5	0.03	✓	✗
	4/5	0.19	✗	✗
Six	6/6	0.02	✓	✗
	5/6	0.12	✗	✗
Seven	7/7	0.008	✓	✗
	6/7	0.06	✗	✗
Eight	8/8	0.004	✓	✓
	7/8	0.04	✓	✓

A score of 5/5 does meet the criteria of $p < 0.05$, however 4/5 does not meet this criteria therefore 5 trials does not allow for any lapses in concentration. The same is true of 6 and 7 trials.

If the participant scored ≥7/8, the criteria of $p < 0.05$ is met, and so 8 trials is the minimum number of trials needed per interval, in order to allow one lapse of concentration. Allowing two lapses of concentration may be considered even better, however a compromise had to be made between this and allowing enough interval sizes plus enough different stimuli in order to produce a wide ranging test in these other dimensions.

Due to the development of the PCT from a 3 note oddball procedure to a 4 AFC contour identification task, each interval would then be presented 4 times. As each contour and interval were to be presented 8 times, this meant that each interval across the 4 contours was presented 32 times. The binomial calculator was used to calculate the number of trials needed to keep the probability of success due to chance below 5%, with chance level = 50%.

Probability of scoring 32/32 by chance = <0.000001

≥31/32 = <0.000001

≥30/32 = <0.000001

≥29/32 = 0.0000013

≥28/32 = 0.0000097

≥27/32 = 0.000057

≥26/32 = 0.00027

≥25/32 = 0.001

≥24/32 = 0.004

≥23/32 = 0.01

≥22/32 = 0.025

≥21/32 = 0.055 (e.g. higher than the alpha of 0.05)

Therefore a score of 22 or better out of 32 was the cut-off point that defined success on a given interval size. This is based on the assumption that each CI user can hear all 4 contours equally. Each trial and response is recorded in the results file, and so if a participant was not able to do this, this information could be discovered, however the summary of scores at the bottom of each condition (e.g. F4 sine) does not currently include a breakdown of scores into the 4 contours, and this would be a beneficial feature for future updates.

There are 6 conditions (F4 sine, F4 complex, F4 piano, F5 sine, F5 complex and F5 piano) and 6 intervals (0.5, 1, 3, 5, 7 and 9 semitones). With 32 trials per interval, this means that for each condition there are 192 trials and for all 6 conditions, there are 1,152 trials.

6.5.5 Interval size

The literature suggests that CI users have the ability to pitch discriminate or pitch rank 1 semitone or less (Gfeller *et al.*, 2002; Nimmons *et al.*, 2008; van Besouw and Grasmeder, 2011; Maarefvand, Marozeau and Blamey, 2013) to 6, 8, 12 or 24 semitones (Gfeller *et al.*, 2002; Jung *et al.*, 2009; Kang *et al.*, 2009; Drennan *et al.*, 2010). Given that CI users may show non-monotonic

psychometric functions, initial plans were to assess a 'wide' range of intervals using the MCS, so that the psychometric function could be estimated. Originally the range was 1-12 semitones, however as comparison of octaves can cause 'octave confusion' and not many CI users performed at ceiling on the UW CAMP, the interval 12 semitones was removed.

Testing every semitone would have led to a very time consuming test and so every other semitone was included, e.g. 1, 3, 5, 7, 9 and 11 semitones. Intervals of smaller than 1 semitone were also considered very important, given the ceiling effects seen in tests that use 1 semitone (Nimmons *et al.*, 2008) and evidence of smaller discrimination thresholds (van Besouw and Grasmeder 2011) and the need to measure future improvement, hence the inclusion of the 0.5 semitone interval. Pilot testing on 2 CI users demonstrated very good ability at 11 semitones, comparable to 9 semitones and so 11 semitones was removed.

6.5.6 Stimuli choices

The inclusion of different frequencies and timbres was important to try to assess across a range of representative stimuli, and not just to assume that ability in one area could be generalised to another. Nimmons *et al.* (2008) chose the notes for the UW CAMP of C4, E4 and G4 because they were in the middle C (C4, 262 Hz) octave, and nursery rhymes are typically presented to children in this octave, plus many instruments are able to play these notes. This same reasoning was used for the PCT, plus this enabled similar octave comparison with the UW CAMP, the MedEl MuSIC Test, the MCI and the SOECIC MTB PDT. The notes F4 and F5 were chosen. The note F5 was chosen because it was one octave above F4 in order to compare identical parameters of the PCT with a higher base frequency. The three timbres (sine, complex and piano) were chosen to represent different harmonic complexities, in order to investigate whether harmonics helped or hindered pitch perception in CI users and NHL.

The complex tones comprised three harmonics: F0, 2F0 at -3 dB and 3F0 at -6 dB as in the MCI test (Galvin *et al.* 2007). The piano tones were generated using the Logic Pro EXS24 sampler. The duration of each pure and complex tone in the PCT was 500 ms with 30 ms linear onset and offset ramps, and the piano tones rung out within 500 ms. Each triplet had 300 ms of silence between notes.

6.6 Pilot testing

A microtonal version (the 'PCTm') was also developed so that the test methodology could be assessed using NHL to establish relationships with similar pitch tests in the NHL population, and to assess reliability on retest for this population. Initially, this test used the intervals 0.05, 0.10, 0.25 and 0.50 semitones, using F4 and F5, and sine, complex and piano timbres. Pilot testing with 10 NHL established that 0.25 and 0.50 were essentially always correct for both discrimination and ranking and so to avoid ceiling effects at those intervals and to enable more information at smaller intervals, and to create a more equal spaced representative psychometric function, the intervals were altered to 0.05, 0.10, 0.15, 0.20 and 0.25 semitones.

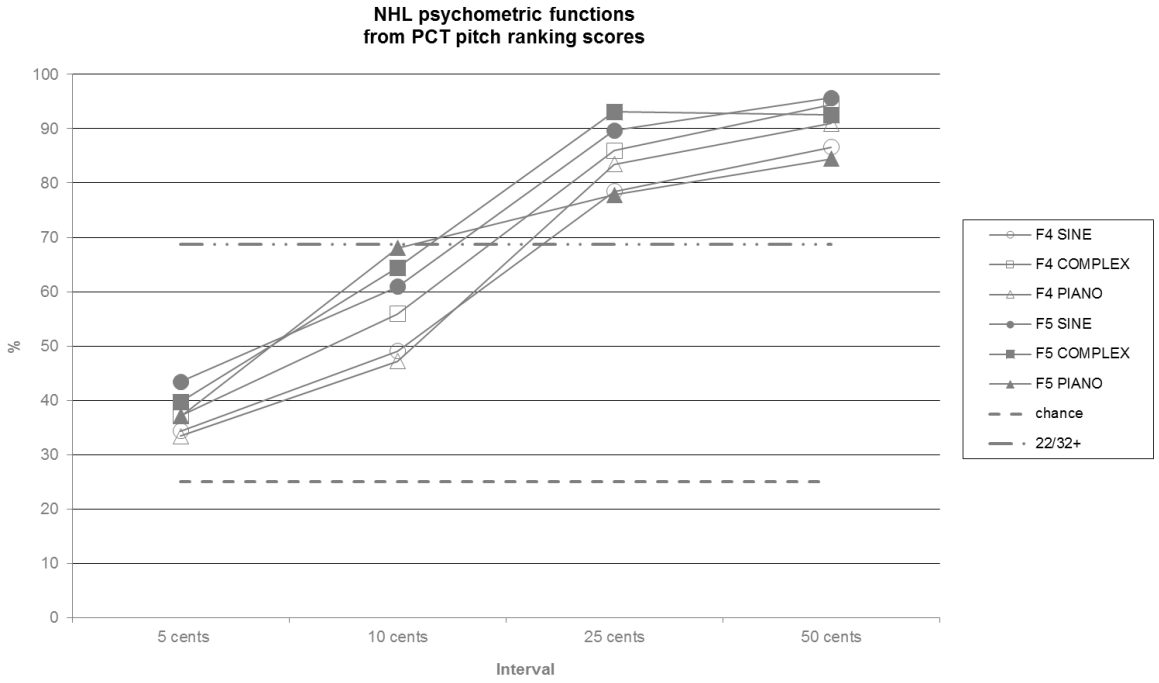


Figure 6.4 Estimated psychometric functions for 10 NHL using the initial version of the PCTm. Intervals of 0.05, 0.10, 0.25 and 0.50 semitone. Upper dotted line represents a score of 22 or more, which is the point at which chance of that score happening is less than 5%. Lower dotted line represents 25% chance level.

Pilot testing with two CI users demonstrated that 11 semitones was typically perceived very well both in terms of discrimination and ranking and so was removed, in order to reduce test time resulting in a range of 0.5, 1, 3, 5, 7 and 9 semitones – this reduced trials per condition to 192, and took approximately 20 minutes.

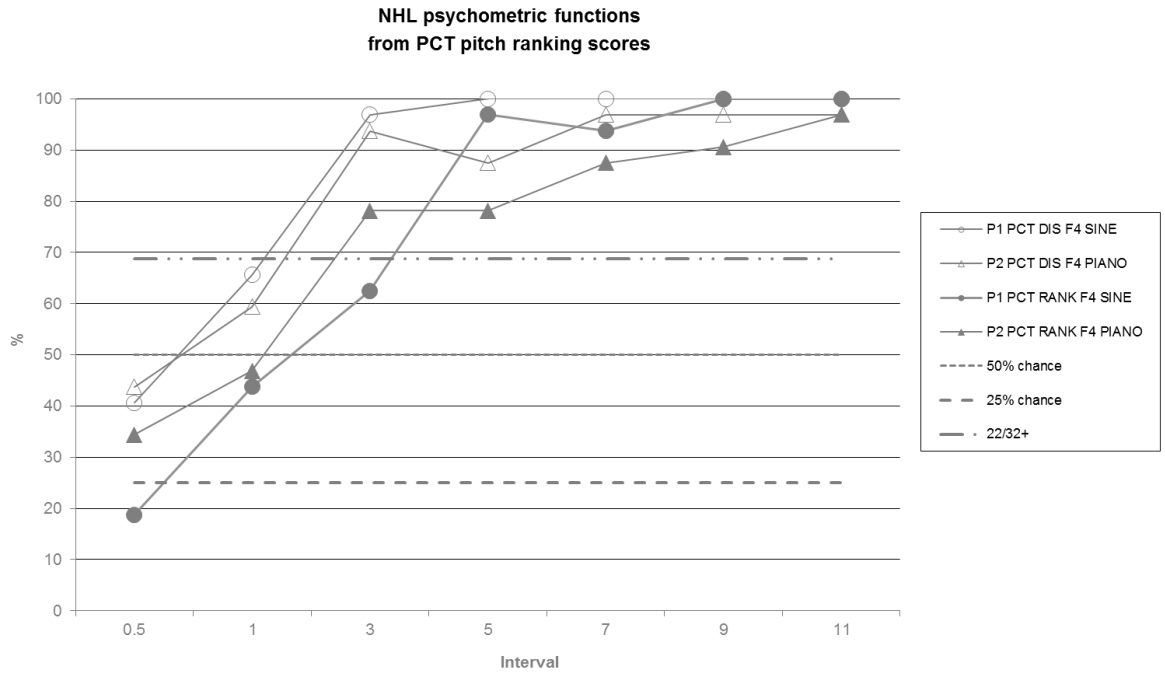


Figure 6.5 Estimated psychometric functions for 2 CI users using the initial version of the PCT. Intervals of 0.5, 1, 3, 5, 7, 9 and 11 semitones. Upper dotted line represents a score of 22 or more, which is the point at which chance of that score happening are less than 5%. Middle dotted line represents 50% chance level and lower dotted line represents 25% chance level.

6.7 Summary

In summary, the PCT:

- tests pitch discrimination and pitch ranking simultaneously, meaning that it can test people who may not be able to pitch rank, but can successfully discriminate tones
- can test pitch discrimination and ranking ability at 0.5, 1, 3, 5, 7 and 9 semitones
- uses graphical images and avoids use of the words 'higher' or 'lower' – similar to the MCI
- avoids the use of adaptive measures, and the use of the MCS means that the psychometric function can be estimated
- has sensitivity of around 2 semitones across the test, but this should be improved upon in future versions

- presents three timbres (a pure tone, a three harmonic synthetic complex tone, and a synthetic piano tone), with two base frequencies (F4, 349.23 Hz and F5, 698.46 Hz)
- presents a similar complexity of contour across all four contours which is an improvement on the MCI

Chapter 7 Experiment 2: Evaluating the Pitch Contour Test

7.1 Introduction

In reviewing existing measures of pitch perception, several issues were raised, which are presented in Table 5.11. The main issues that were raised were inappropriate difficulty levels for CI users, lack of evidence of test retest reliability with CI users and the use of adaptive procedures with CI users. The test with the best properties was the MCI, which showed minimal floor and ceiling effects, showed very good reliability on retest and used the MCS. There were some issues with the MCI including different levels of complexity in the contours, and a lack of clarity in terms of measuring individual pitch intervals. This lead to the design of the PCT, which aimed to create a simpler version of the MCI, using only 4 similarly complex contours, which tested using the MCS, tested a range of intervals from 0.5 – 9 semitones using different stimuli.

This chapter will present the evaluation of the PCT and its comparison with the MCI, the UW CAMP and the SOECIC MTB PDT, with CI users and NHL. The MCI was chosen for comparison due to its superior performance in Experiment 1, and because the PCT shares a similar design, and due to its wide usage and strong presence in the literature. The UW CAMP was also chosen to compare to the PCT due to its wide usage and strong presence in the literature. Both the MCI and the PCT were adapted for use with NHL by using microtonal notes. The SOECIC MTB PDT was also included in the NHL comparison group as it was the only test from Experiment 1 with suitable difficulty for NHL.

7.1.1 Aims

There were two main aims to Experiment 2:

1. To establish whether the PCT is a suitable method of pitch perception for CI users and to determine whether it can be considered an improvement on existing tests
2. To determine the extent to which CI users' psychometric functions are non-monotonic

7.1.2 Objectives

The objectives for Experiment 2 are based on the evaluative criteria described in section 5.1.2:

Appropriateness of content	Is the difficulty level suitable? Are floor and ceiling effects avoided?
Construct validity	Does the test provide a suitable number of repeats to ensure statistical confidence in the results? Is the methodology used to calculate the score suitable for use with CI users?
Criterion (concurrent) validity	Do test results correlate with theoretically similar tests, taken at the same time? Are differences seen between musicians and non-musicians?
Reliability	Does the test produce a similarly repeatable score on retest?

7.1.3 Research questions

1. Does the PCT provide enough repeats to give statistical confidence in the final result?
2. Does the PCT provide a suitable level of difficulty for test users?
3. Does the PCT demonstrate reliability on retest?
4. How do the PCT results compare with existing literature and similar tests?
5. Does the PCT demonstrate sensitivity to musicianship?
6. Are CI users' psychometric functions monotonic?
7. Are PCT results affected by stimulus type?

7.1.4 Hypotheses

7 The PCT will provide enough trials to give statistical confidence in the final result

8 The PCT will provide a suitable level of difficulty for CI users

9 The PCT will be reliable on retest

10 The PCT may show some similar results between existing tests although the differences in the methodologies may mean that this will not be seen for all tests.

11 The PCT will show sensitivity to musicianship

12 At least some of the CI users' psychometric functions will be non-monotonic

7.2 Methods

7.2.1 Materials - the pitch tests

In addition to the PCT, a subset of the tests used in Experiment 1 were also used in this chapter:

1 The MCI test (Galvin, Fu and Nogaki, 2007)
2 The UW CAMP: pitch test (Nimmons *et al.*, 2008)
3 The SOECIC Music Test Battery: pitch identity test (van Besouw, 2010)

Details of the tests are given in section 5.3.1. In addition to the above tests, the MCI was also adapted for use with NHL by using 'microtonal' (e.g. pitch intervals of less than 1 semitone) stimuli in place of the original stimuli (the 'MCIm'), and the PCT and a 'microtonal' version of the PCT, the PCTm was also used. Details of these additional tests are given in Table 7.1.

Table 7.1: Additional tests used in Experiment 2

Test	Stimuli & Range	Interval size	Test type, number of trials, chance level
PCT	sine tone (generated in Adobe Audition) 3 harmonic complex tone (generated in Adobe Audition):F0, 2F0 at -3 dB and 3F0 at -6 dB piano tone generated using Logic Pro EXS24 sampler	0.5, 1, 3, 5, 7, 9 semitones Frequencies of notes (F4 and F5) calculated using A=440 Hz and semitone difference of 12^{th} root of 2	MCS, 32 trials Discrimination chance : 50% Ranking chance : 25%
PCT Microtonal (PCTm)	sine tone (generated in Adobe Audition) 3 harmonic complex tone (generated in Adobe Audition):F0, 2F0 at -3 dB and 3F0 at -6 dB piano tone generated using Logic Pro EXS24 sampler	0.05, 0.10, 0.15, 0.20, 0.25 semitones Frequencies of notes calculated using A=440 Hz and 1 cent (x5 = 5 cents) difference of 1200^{th} root of 2	MCS, 32 trials Discrimination chance : 50% Ranking chance : 25%
MCI Microtonal (MCIm) Adapted from Galvin, Fu and Nogaki (2007)	Synthesized complex tone, 3 harmonics, as in Table 5.2 A4 = 440 Hz A4+5(cents) = 441.28 Hz A4+10 = 442.56 Hz A4+15 = 443.84 Hz A4+20 = 445.13 Hz A4+25 = 446.42 Hz A4+30 = 447.72 Hz A4+35 = 449.02 Hz A4+40 = 450.32 Hz	0.05 semitone (5 cents) interval tests 0.10 – 0.20 semitones 0.10 (10 cents) semitone interval tests 0.20 – 0.40 semitones Frequencies of notes calculated using A=440 Hz and 1 cent (x5 = 5 cents) difference of 1200^{th} root of 2	5 x 27 9AFC contour discrimination Chance = 11%

7.2.2 Equipment

Data collection was conducted in an acoustically treated room at the ISVR with ambient noise levels of <35 dB(A). All tests were run using a Dell Latitude E6400 laptop running Windows XP, an external Behringer UCA202 soundcard and Behringer Truth B2031P loudspeaker which was positioned 150cm from the position of the listener's head, with the tweeter at ear level. A flat screen monitor was positioned in front and slightly to one side of the participant so that they could see the graphical user interface for each test and respond using a computer mouse.

7.2.3 Calibration

As in Experiment 1, stimuli for all experiments were presented in close to 'free field' conditions. The aim was to present all stimuli at the participant's ear at 65 dB(A). This was checked in the absence of the participant, using a hand held Bruel and Kjaer Type 2235 Precision SLM mounted on a tripod at the position of the ear, 1.5m away from the loudspeaker, at tweeter level. Both the UW CAMP and the SOECIC MTB PDT provided their own calibration tone, however prior to testing participants, the tests were run by the experimenter to check that the wide range of frequencies fell between the tolerated levels of 60-70dB(A).

For the two microtonal tests, the MCIm and the PCTm, the contours (MCIm) and the triplets (PCTm) were created so that each note of the contour and triplet was adjusted in loudness to be at 65dB(A), at the position of the ear, using the SLM set up described above, prior to being put together to make the contour or the triplet.

7.2.4 Ethical approval

Ethical approval was obtained from the National Research Ethics Service (NRES) reference number 11/SC/0263 and from the UOS Institute of Sound and Vibration Research Safety and Ethics Committee, references 1135 and the UOS's Research Governance Office, reference 7511. Informed written consent was taken from all participants prior to proceeding.

7.2.5 Participants

Twenty-three NHL and 22 CI users took part in this experiment.

Study 4 - NHL

The 23 NHL (11 female, 12 male, 14 musicians[4], 9 non-musicians, mean age 25 years, ±4 SD) were recruited via opportunity sampling from the University of Southampton, were aged between 18 and 32, and all had hearing thresholds in quiet of 20 dB HL or better, and reported no amusia or tone deafness, prior to taking part in the study.

Study 5 – CI

The 22 CI users (14 female, 8 male, 3 musicians 19 non-musicians, mean age 64 years, ±12 SD) were recruited by two methods: invitations were sent to CI users who were under the care of the University of Southampton Auditory Implant Service (USAIS, formerly SOECIC); and invitations were sent to CI users from around the UK via the National Cochlear Implant User Association (NCIUA) e-mail newsletter, and the same invitation was sent by e-mail to specific NCIUA local support groups: the Gloucestershire Deafened and Cochlear Implant Support Group; the Home Counties Cochlear Implant Group; the Oxford Cochlear Implant Support Group; the South of England Cochlear Implant Users Group and the West of England Support Group. These groups were chosen because they were within reasonable travelling distance of the University of Southampton. Testing took place between June and September 2013.

[4]*Musician* defined by self-report: if any participant confirmed that they held any musical qualification, or considered themselves to be a musician, they were considered to be a musician for this thesis.

Table 7.2 Study 5 CI user demographics

ID	Age (years)	Sex	Pre/post lingual deafness	CI manufacturer	Listening mode	Duration post initial tuning (months)	Prior music training
CI1	62	F	Post	Advanced Bionics	Unilateral	66	School
CI2	88	F	Post	Cochlear	Unilateral	60	None
CI3	73	F	Post	Cochlear	Unilateral	204	School
CI4	49	F	Post	Advanced Bionics	Unilateral	54	School
CI5	70	F	Post	Cochlear	Unilateral	60	School
CI6	65	M	Post	Cochlear	Unilateral	216	None
CI7	45	F	Pre	MED-EL	Unilateral	18	None
CI8	40	M	Pre	Advanced Bionics	Unilateral	24	None
CI9	66	M	Post	MED-EL	Unilateral	18	None
CI10	53	F	Post	MED-EL	Unilateral	3	None
CI11	66	F	Post	Advanced Bionics	Unilateral	123	School
CI12	68	F	Post	MED-EL	Unilateral	120	School
CI13	63	M	Post	Advanced Bionics	Unilateral	84	Musician
CI14	55	F	Post	MED-EL	Unilateral	18	None
CI15	66	F	Post	Cochlear	Unilateral	108	Musician & school
CI16	70	M	Post	Advanced Bionics	Unilateral	12	School
CI17	47	F	Post	Advanced Bionics	Unilateral	18	Musician & school
CI18	69	M	Post	MED-EL	Bilateral	180	School
CI19	74	M	Post	Advanced Bionics	Unilateral	36	None
CI20	62	F	Post	Cochlear	Unilateral	11	None
CI21	81	M	Post	Cochlear	Unilateral	12	None
CI22	68	F	Post	MED-EL	U	18	None

The general inclusion criteria for the study was that potential recruits had to be adults with one or two CI and be resident in mainland UK. USAIS is associated with a large number of research projects and as such has methods in place to ensure that patients are not over invited to participate in research. This meant that certain extra exclusion criteria were included, for recruitment purposes only, for patients recruited through USAIS, e.g. patients that had a switch on date of less than 12 months prior to recruitment were not included. Patients that were considered by the research coordinator to be unlikely to cope well with the test environment were not invited.

Patients with severe visual impairment were not included, as the visual nature of the touch screen set up meant that this was not ideal. All CI users were offered a payment of £20 for their time, and travel and accommodation costs were reimbursed.

7.2.6 Procedure

NHL

The NHL participants attended for two sessions. Hearing thresholds were measured to ensure that they were at 20 dB HL or better for octave frequencies between 250 Hz and 8000 Hz, if this had not been completed within the last 6 weeks, and no changes in hearing were reported. Each participant was invited to read the participant information sheet and sign the consent form, plus answer a few questions about their musicianship status.

Each participant took part in three tests of pitch perception, the PCTm, the SOECIC MTB PDT and the MCIm. Order of testing was randomised across participants, using a Latin Square design, in order to minimise order effects and learning effects (as described in section 5.2.2). This was used to determine the order of tests (PCTm, CAMP and MCIm) and the order of block presentations within the PCT (sine, complex and piano) for each participant. The interval sizes of the MCIm however, were always presented with the larger interval first (e.g. 0.1 semitones) as the MCI is a difficult test and it was felt that starting with the smallest interval might be disheartening for participants.

Levels were checked at the position of the participant ear using a Bruel and Kjaer Type 2235 Precision SLM. Participants were seated at 1.5m in front of the speaker, with their eyelevel in line with the tweeter. During testing, participants responded using a mouse on their lap, and interacted with the graphical user interface via a computer screen positioned to the left of the loudspeaker.

CI users

The CI participants also attended for two sessions over 2 days (except for 5 CI users who preferred to attend for just one day and complete all session within one day (CI 4, 6, 10, 14 and 19) and CI 8 who was not able to attend for the second day. The time between session 1 and session 2 ranged between 1 and 21 days (mean 5.5 days). Each participant was invited to read the participant information sheet and sign the consent form. Participants were asked to complete a few questions about their musicianship status. Participants were asked to use their ‘everyday’ program on their CI, and if they wore a contralateral HA, were asked to continue to wear it, but switch it off, during testing. The hearing aid was checked by the experimenter to make sure that it was switched off.

Participants took part in three tests of pitch perception: the MCI, the UW CAMP and the PCT. The order of these tests was randomised by use of Latin square (as described in section 5.2.2), however the interval sizes of the MCI were always presented in the same order e.g. 5 semitones interval was presented first, followed by 4, 3, 2 and 1. Prior to each pitch test, instructions were given and the researcher clarified with each participant that they understood the task. The structure of testing days is shown below:

Day 1:	Morning session	T1 PCT F4 (stimuli order random), UW CAMP, MCI (test order random)
	Afternoon session	T2 PCT F4 (order as above), UW CAMP, MCI (order as above). Each participant had a lunch break of one hour between the end of session 1 and the start of session 2.
Day 2:	Morning session	T1 PCT F5 (stimuli order random)
	Afternoon session	T2 PCT F5 (order as above)

The reason for the above structure was because not all participants could attend for two days, and so priority was given to getting repeat testing for the PCT (F4) plus the comparison tests, and so getting initial and repeat data for F5 was a bonus in those who were able to attend for both days (16 of 22 CI users). At the end of each test, participants had their results explained to them.

7.3 NHL Results

This section presents the results of the PCTm test, and compares them with the results from the SOECIC MTB PDT and the MCIm tests, using a group of 23 NHL. Data from T2 was only used for the reliability analysis and the estimation of psychometric functions.

Initially, the SOECIC MTB PDT, being the only adaptive test in this study, was analysed to determine whether the number of trials was sufficient to minimise the effects of chance on results. The effect of floor and ceiling effects were investigated for all 3 tests, psychometric functions were estimated for NHL using the PCTm, PCTm scores were calculated, and data loss issues were described. Results from the PCTm and SOECIC MTB PDT were compared. The effect of musicianship was

analysed, test retest reliability was calculated and the effects of stimuli were investigated. Medians and IQRs are presented in section 7.3.6 (rather than earlier) because earlier sections provide details about how the PCTm can be scored to allow it to be compared with other tests.

7.3.1 The role of chance in the SOECIC MTB PDT

As described in section 5.4.1, adaptive methods result in differing numbers of trials per run. These trial numbers were considered in respect to the binomial distribution and the role of chance. These calculations were not undertaken for every participant, however two examples were taken for each adaptive test, a good performer and a poor performer.

SOECIC MTB PDT: 3 AFC, chance = 33%, 2 down 1 up, no repeats

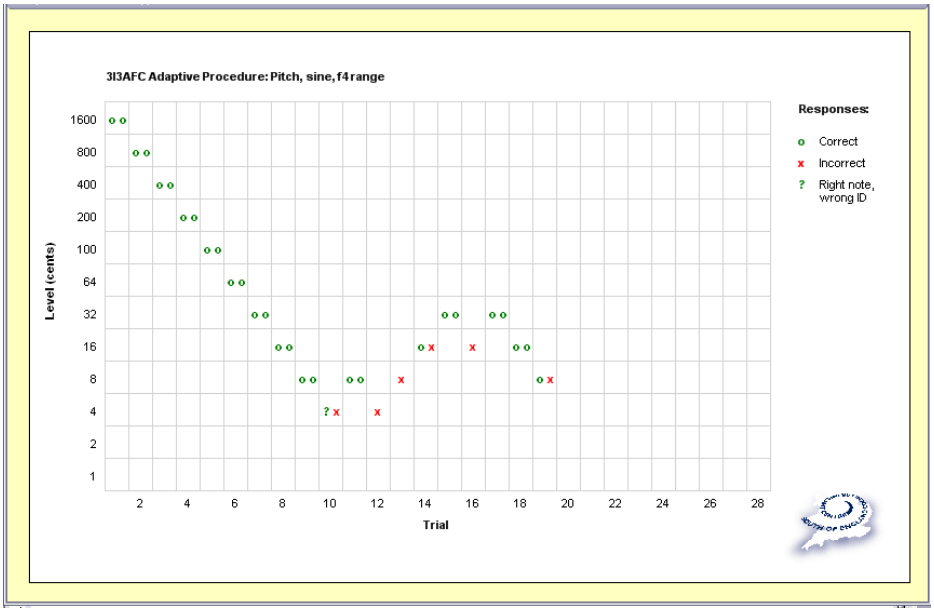


Figure 7.1 Example of a good performer (NHL 8) using the SOECIC MTB PDT, showing the adaptive staircase. Final score was 18.4 cents (0.184 semitones). There were sufficient trials within this test (and with this level of performance) to ensure that the final score had $p < 0.05$.

In the example above, the final score was 18.4 cents (0.184 semitones). The final score of the SOECIC MTB PDT is calculated by averaging the final 5 of the 7 reversals. Figure 7.1 shows the final 5 reversals (in reverse order) as 8, 32, 16, 32 and 4 cents, and their average is 18.4 cents.

Using the binomial calculator, assuming a probability of success on a single trial being 0.333, the following numbers of successes are needed in order to keep the likelihood of success due to chance below 0.05:

Successes/trials	*p* of scoring that *due to chance*
Score = 3/3	0.04
Score = 4/4	0.01
Score = 4/5	0.05
Score = 5/6	0.02
Score = 5/7	0.05

Below is a breakdown of NHL 8's performance, for each interval presented, the number of successes and the number of trials. Scores of 2/2 from the intervals 1600 – 64 cents are deliberately not included in the presentation below. Asterisks show that enough trials were presented for 32, 16 and 8 cents (indicated by *) to keep $p < 0.05$.

32 cents: 6/6 successful*

16 cents: 5/7 successful*

8 cents: 5/7 successful*

4 cents: 1/3 successful

The results from this test demonstrate that NHL 8 could discriminate 32, 16 and 8 cents, as the numbers of correct scores indicated that the chance of this listener getting these scores by chance was less than 5%. Therefore, the final test score for NHL 8 of 18.4 cents was close to these scores, particularly close to the tested interval of 16 cents. However it could be argued that a score of closer to 8 cents might more accurately reflect NHL 8's performance.

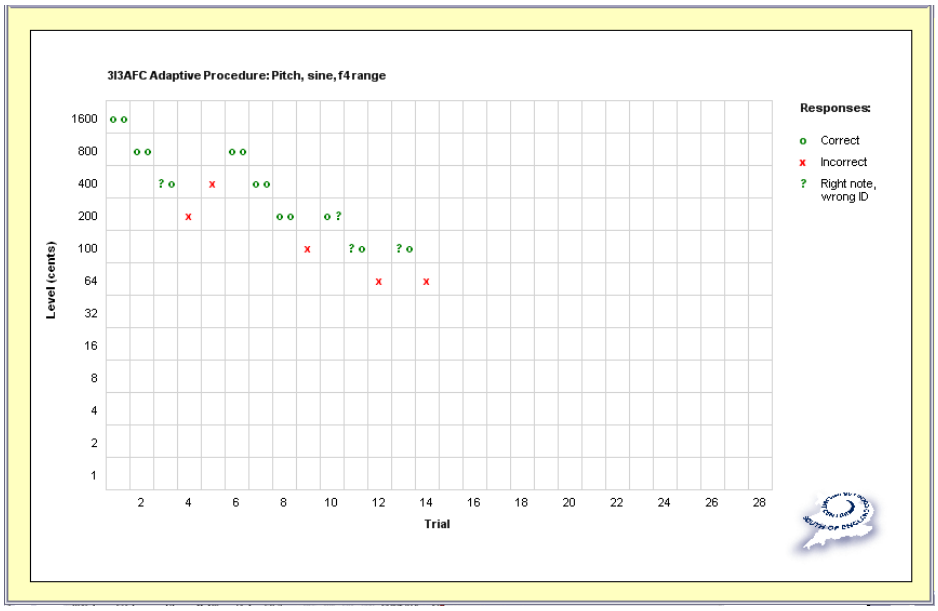


Figure.7.2 Example of a poor performer (NHL 14) using the SOECIC MTB PDT, showing the adaptive staircase. Final score was 105.6 cents (1.056 semitones). There were sufficient trials within this test (and with this level of performance) to ensure that the final score had $p < 0.05$.

In the example above, the final score was 105.6 cents (1.056 semitones). Figure.7.2 shows the final 5 reversals (in reverse order) as 64, 100, 64, 200 and 100 cents, and their average is 105.6 cents.

Below is a breakdown of NHL 14's performance, for each interval presented, the number of successes and the number of trials.

1600 cents: 2/2 successful

800 cents: 4/4 successful*

400 cents: 4/5 successful*

200 cents: 4/5 successful*

100 cents: 4/5 successful*

64 cents: 0/2 successful

This indicates that enough trials were presented for 800, 400, 200 and 100 cents (indicated by *) in order to keep $p < 0.05$. The final score was close to the interval 100 cents and so the likelihood that this was due to chance is small.

7.3.2 Floor and ceiling effects

As in Chapter 5, ceiling effects were defined as the best possible score that is achievable per test, and floor effects were defined as the lowest possible score (for adaptive methods) or scores at or below the chance level per test. For the SOECIC MTB PDT, this was 16 semitones and 0.01 semitones, and for the MCI this was 100% and 11%. Floor and ceiling effects were analysed using the raw data from the PCTm and as such the definition was adapted in order to encompass data from all intervals. Rather than define ceiling as 100% for each interval, ≥80% was used instead, as this would otherwise seemingly give an unfair advantage to the PCT, plus the ceiling effects that were seen using this definition typically had a score of ≥80% for the smallest interval and 100% for all other intervals (for PCTm). Floor effects were defined as ~≤ chance level, which was 50% for PCTm discrimination and 25% for PCTm ranking. These were reported for an individual if it occurred for at least 1/6 conditions. Floor and ceiling effects were reported for T1 data only, as this reflects the way these tests might be performed clinically.

PCTm

PCTm discrimination: Ceiling effects were seen for 3/23 NHL: NHL 4 (F4 sine), NHL 9 (F4 complex, F4 sine, F5 sine, F5 complex), NHL 23 (F4 complex, F4 sine, F5 complex). Floor effects (~50% for all intervals) were seen for 2/23 NHL: NHL 14 (F4 sine), NHL 17 (F4 sine, F4 piano, F5 piano). PCT ranking: Ceiling effects (>80% for all intervals) were seen for 2/23 NHL: NHL 9 (F4 complex), NHL 23 (F4 complex, F5 complex). Floor effects (~25% for all intervals) were not seen. Examples are shown in Figure 7.3.

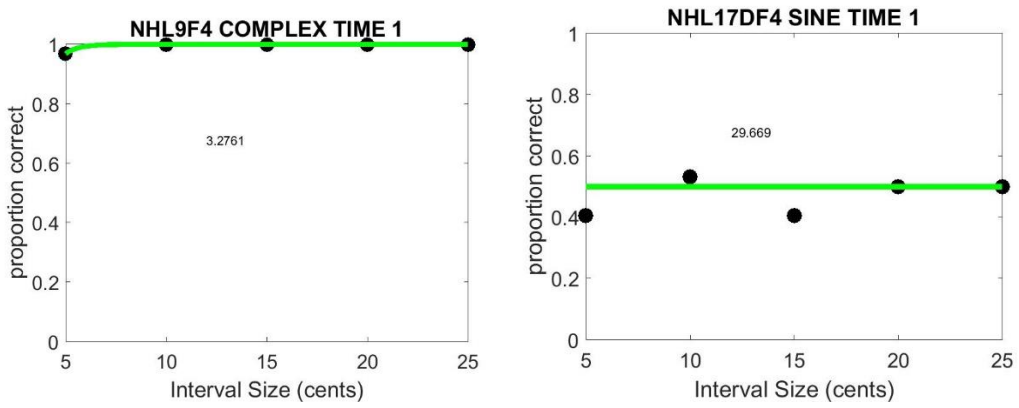


Figure 7.3 Examples of ceiling and floor effects in the PCTm, with NHL. Ceiling: PCTm ranking, NHL 9, F4 complex, ≥80%. Floor: PCTm discrimination, NHL 17, F4 sine, ~≤50%. NHL9F4 = NHL 9, with note F4, pitch ranking task. NHL17DF4 = NHL 17, with note F4, with pitch discrimination task.

SOECIC MTB PDT

Ceiling and floor effects had to be defined differently for the SOECIC MTB PDT as scoring over 80% or at chance did not apply to this test because this test did not provide percentage results, rather it determined thresholds in terms of cents and semitones. Instead, anyone not succeeding at 16 semitones would be considered to be performing at floor level, and anyone succeeding at 0.01 semitones would be considered to be performing at ceiling. This was not seen within the group of NHL users. The range of scores from the 23 NHL from T1 was from 0.06 to 1.2 semitones.

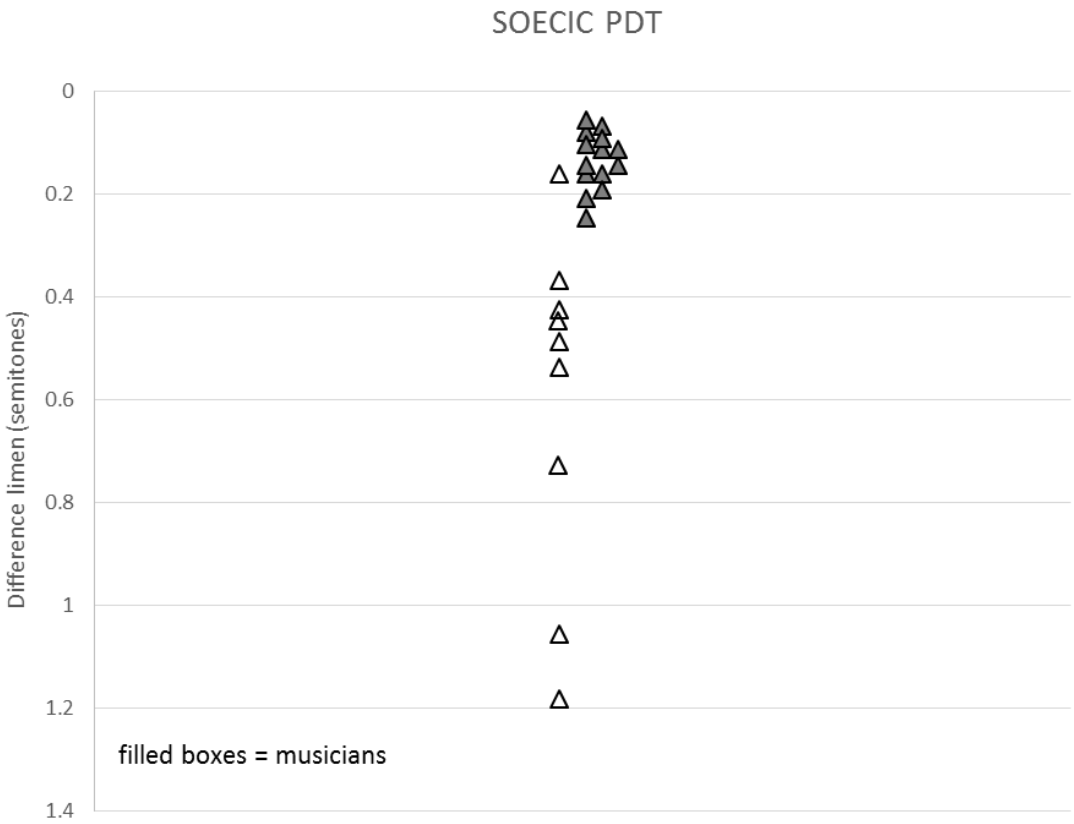


Figure 7.4 NHL SOECIC MTB PDT scores showing range of scores and effect of musicianship, with musicians demonstrating much smaller difference limens compared to non-musicians.

MClm

Ceiling effects (100%) were seen for 0.1 semitones only, for 3/23 NHL (NHL 8, 9 and 23). Floor effects (≤11%) were seen for 1/23 (NHL 17) for 0.1 semitone and 3/23 for 0.05 semitone (NHL 11, 15, 18).

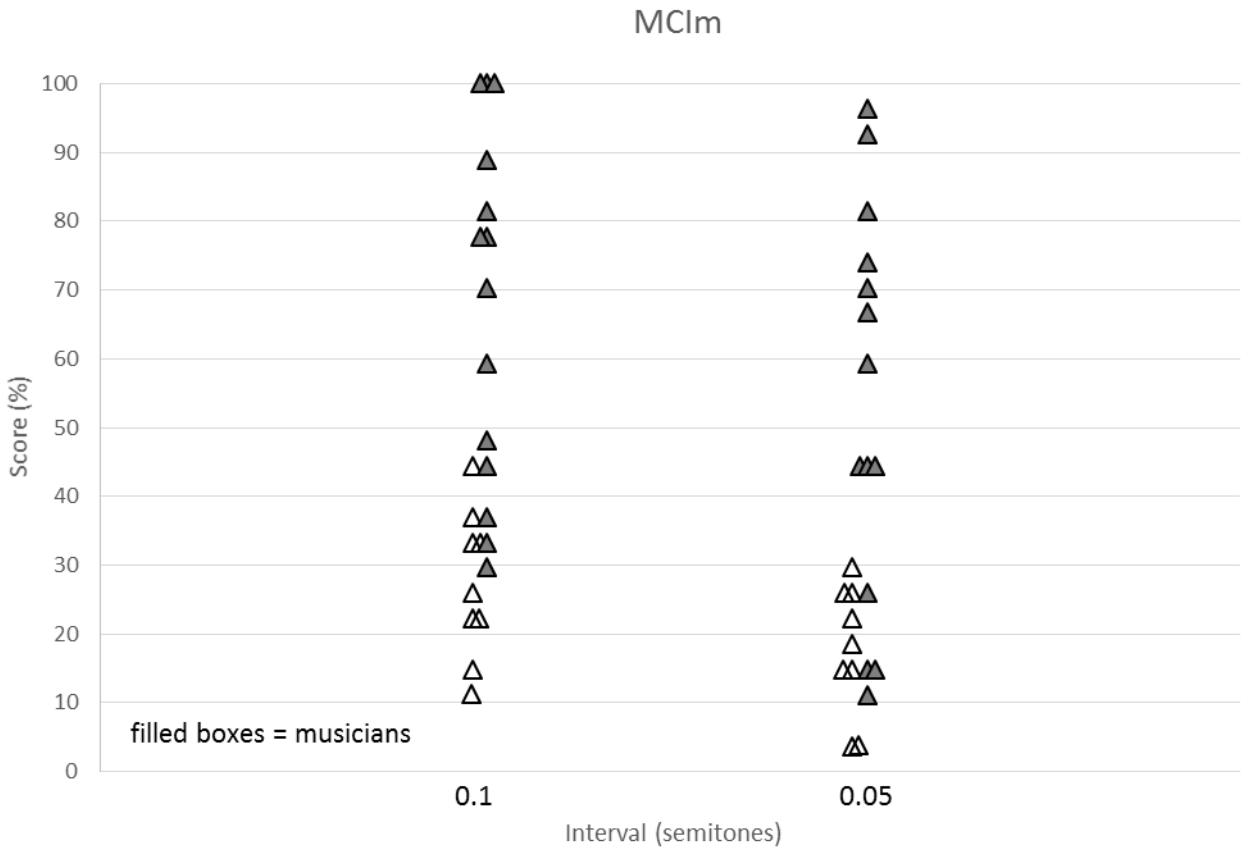


Figure 7.5 NHL MCIm scores showing range of scores and effect of musicianship

7.3.3 Psychometric functions for NHL

The output of the PCT and the PCTm is presented within a results output file and provides details of how many correct answers were obtained for pitch discrimination and pitch ranking for each of the 3 timbres at the 2 frequencies. A typical example of this can be seen in Table 7.3.

Table 7.3 Example of summary output 'results file' from the PCT

Instrument: Sine (F4)

	Trials	RANK SCORE	RANK %	Trials	DIS SCORE	DIS %
0.5 semitone	32	6	18.75	32	13	40.625
1 semitone	32	14	43.75	32	21	65.625
3 semitones	32	20	62.5	32	31	96.875
5 semitones	32	31	96.875	32	32	100
7 semitones	32	30	93.75	32	32	100
9 semitones	32	32	100	32	32	100

This allowed the psychometric function to be estimated. Psychometric functions were fitted to the data using the MATLAB Palamedes toolbox (Prins and Kingdom, 2009), using a maximum likelihood criterion. The logistic form of the psychometric function was chosen due to its similarities to the normal curve, whereas the Gumbel (which is typically used to model rare events such as peak height of river levels to predict flooding) and Weibull (which is typically used to model time to failure, such as early failure of domestic appliances) curves took a more extreme approach, which was not considered to be appropriate (www.weibull.com/hotwire). A 4% lapse rate, rather than assuming that participants would always achieve 100% at the largest intervals, was chosen to reduce effect of bias when using a maximum likelihood method (Wichmann and Hill, 2001). By not assuming that the participant will always get 100% at the largest intervals (for example if they cough or mishear during testing), the fitting of the maximum likelihood psychometric function is not subject to bias, and Wichmann and Hill demonstrated this in their first 2001 paper.

Each participant (n = 23) attempted 6 conditions at T1 and T2, with the exception of NHL 21 who was unable to return for T2 testing, NHL 7 only completed 10/12 of the conditions due to issues with the sine stimuli and NHL 11 found the task so difficult that he did not wish to continue with the remaining conditions. As such, 256 psychometric functions were generated for the PCTm discrimination test and 256 for the ranking test. Examples of monotonic psychometric functions can be seen below in Figure 7.6.

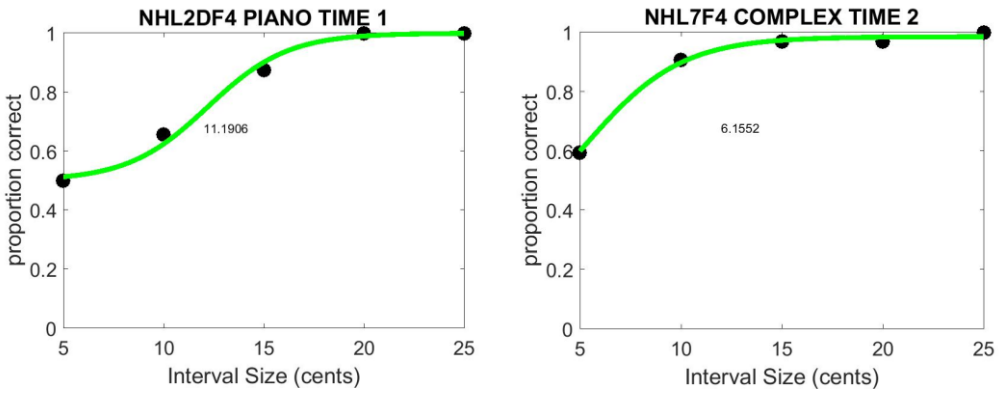


Figure 7.6 Two examples of NHL monotonic psychometric functions. NHL2DF4 = NHL 2, with note F4, pitch discrimination task. NHL7F4 = NHL 7, with note F4, with pitch ranking task.

Non-monotonic functions were defined as scores that did not follow either a positive or a negative trend across the intervals, e.g. often showing better scores at smaller rather than larger intervals. No examples of a reverse function were seen e.g. best performance at smaller intervals and poorest performance at large intervals. Flat scores across intervals were not counted as non-monotonic. Functions were labelled as non-monotonic individually; e.g. a non-monotonic label did not require both T1 and T2 to display non-monotonic features. There were 7 non-monotonic functions seen in the NHL (7/512 = 1.4%), and 2 examples are shown in Figure.7.7 below. All seven non-monotonic functions can be seen in Appendix B. The evidence of non-monotonic psychometric functions is of particular interest in both NHL and CI users as it indicates that an adaptive procedure would be invalid: because monotonicity is assumed within an adaptive staircase methodology (Levitt, 1971) a successful score would lead to the interval size being reduced, which is assumed to be harder. With a non-monotonic psychometric function, this is not the case and therefore the methodology would asymptote to an erroneous value, with no way of the test user knowing that this was the case.

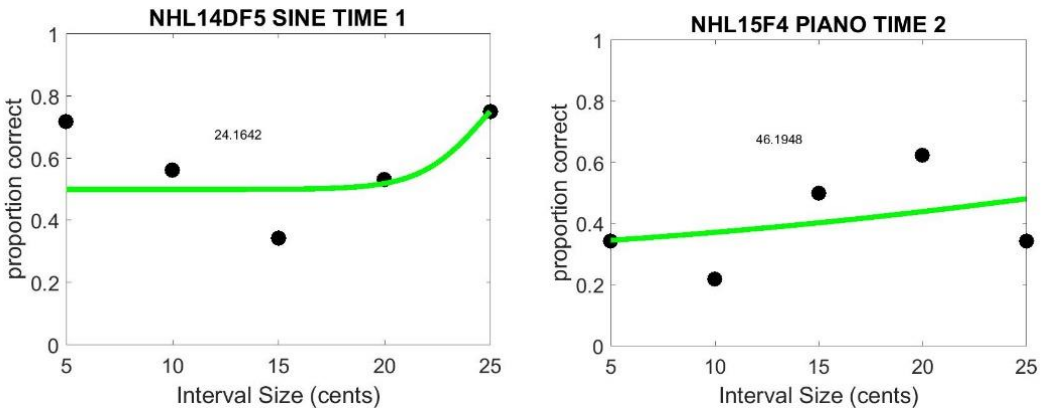


Figure.7.7 Two examples of NHL non-monotonic psychometric functions. NHL14DF5 = NHL 14, with note F5, pitch discrimination task. NHL15F4 = NHL 15, with note F4, with pitch ranking task.

No scores significantly below chance level were seen in the NHL group, indicating that there were no pitch reversals.

7.3.4 Calculation of the PCT Difference limen

One of the design features of the PCT was to use the MCS in order to estimate the psychometric curve rather than assume it is monotonic. The scoring of the PCT as shown in section 7.3.3 above means that an individual may score better for some intervals than for others and the nature of this may not be predictable; it may not be monotonic. In order to provide feedback to participants, a cut off score of 22/32 was used to determine that that interval was 'successfully' discriminated or ranked (as described in Chapter 6). Therefore each interval could be described in a binary way e.g. successful or unsuccessful.

This method of scoring did not allow comparisons with other tests that produce a threshold score, e.g. the MedEl MuSIC Test, the SOECIC MTB PDT or the UW CAMP. In order to make these comparisons, and to statistically analyse the output from the PCTm, a threshold score or 'difference limen' (DL) was calculated. Although this meant that non-monotonic data might be misinterpreted this way, there were only 1.6% of psychometric functions that were non-monotonic within the NHL data.

Success per interval size (e.g. 1, 3, 5 semitones) was determined by the participant scoring 22 or greater out of 32, in order to keep the probability of this score being due to chance lower than the alpha level of 0.05. The same principal was applied to calculating a threshold score using the

MATLAB Palamedes Toolbox. This enabled DLs to be estimated that corresponded to the point on the curve where the participant scored 22/32 (68.75%).

The Palamedes Toolbox enabled a MATLAB file to be used which fitted a psychometric function to the data, using the following parameters:

A maximum likelihood procedure was used, with a logistic function

The threshold point was set as 68.75% which corresponded to a score of 22/32

The lower bound of the curve was set to chance level: for discrimination scoring, this was set to 50%, and for ranking scoring this was set to 25%.

The upper bound of the curve was set to 96%, rather than 100%, to allow for a lapse rate of 4%.

This procedure calculated a DL for each curve to 4 decimal places. The rules regarding whether these DLs are then usable are detailed in Section 7.3.5.

Calculation of the DL raises the same issues as assuming the psychometric curve is monotonic: this ignores the evidence that CI users do not reliably show monotonic psychometric functions. As such, this risks the loss of data accuracy and information regarding the complex nature of the cochlear and electrode relationship and the resulting perception of pitch. For these reasons, the PCT does not automatically calculate the DL as it is not considered helpful for participant feedback.

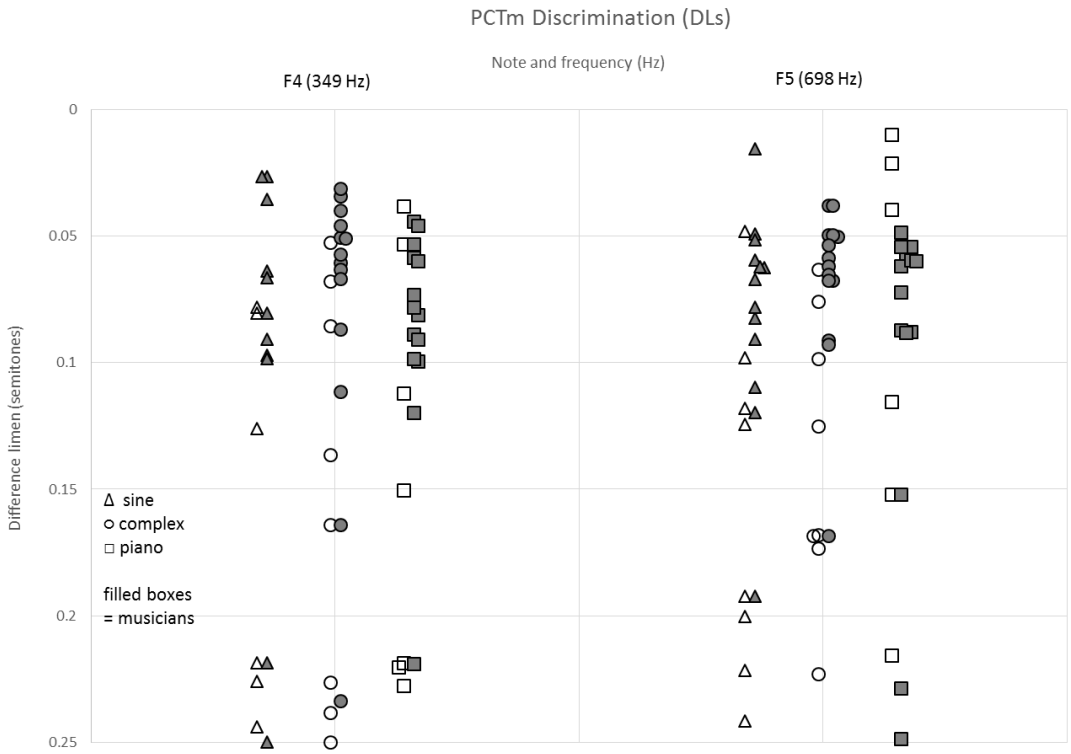


Figure 7.8 PCTm discrimination scores showing effect of musicianship. Filled symbols represent musicianship, and this group were amongst the highest performers (at the top of the figure) however there are several non-musicians who are performing equally well and better than musicians, especially for F5 piano.

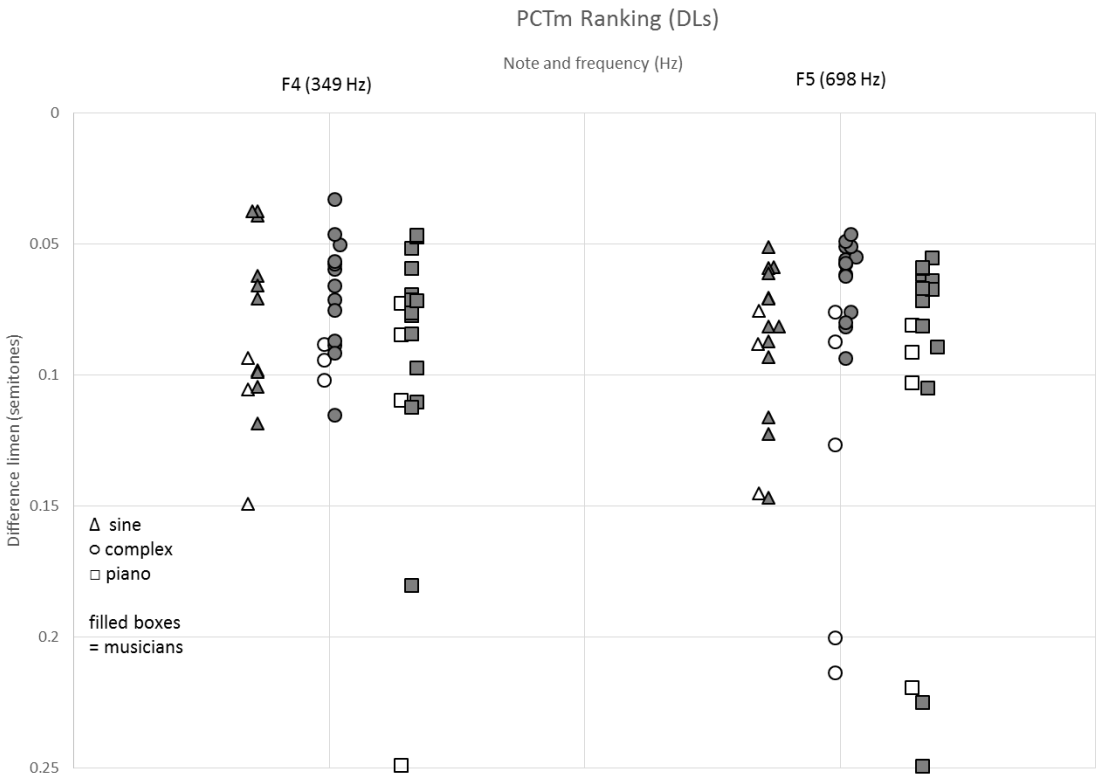


Figure 7.9 PCTm ranking scores showing effect of musicianship

7.3.5 Data rejection due to being outside of the bounds of the PCT

Negative DL values were a problem for the PCT. In 2 cases (within NHL discrimination and ranking combined) the DL was negative, and so these were rejected (Figure 7.10).

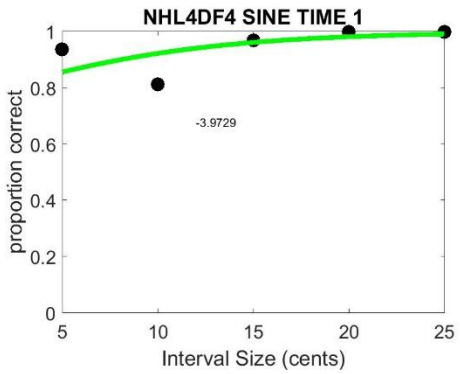


Figure 7.10 Example of negative difference limen. The trend line can be seen exiting the graph at around 0.85, and so the calculated point at which the trend line would hit 0.6875 is -3.97, and a threshold of negative semitones is not possible. This demonstrates a practical issue with the way in which the DL is calculated within the PCT and highlights the problems with the current algorithm for very good performers.

Upper bound rejection

The bounds of the PCTm were 0.05 and 0.25 semitones, and so 0.25 was used as an upper bound, which meant 67 functions were rejected, because the DL was greater than 0.25 (14 from discrimination and 53 from ranking, example given in Figure 7.11).

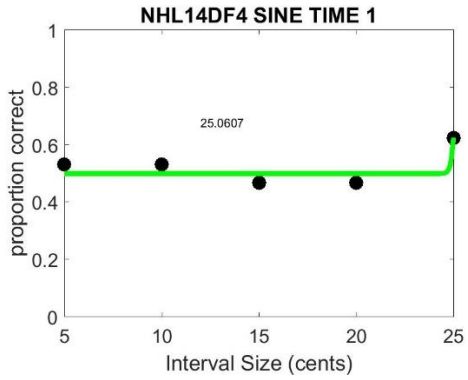


Figure 7.11 Example of DL greater than 0.25 semitones

Lower bound rejection

Each DL below 0.05 semitones was assessed individually for visual goodness of fit, which resulted in 8 being rejected (2 from discrimination and 6 from ranking, example given in Figure 7.12).

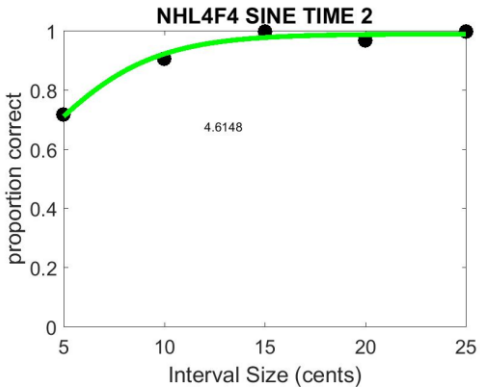


Figure 7.12 Example of DL below 0.05 semitones, e.g. 4.61 cents, or 0.046 semitones.

The total loss of data due to the DL calculation was 17.7%.

7.3.6 Median scores for the PCTm, SOECIC MTB PDT and the MCIm

Descriptive statistics are presented here so that the PCTm DL scores can be compared with the scores from the SOECIC MTB PDT and the MCIm. The Shapiro-Wilk test of normality was used due to its ability to cope with low numbers of *n* (Shapiro and Wilk, 1965). Data distribution was found to be significantly non-normal at the $p < 0.05$ level for all tests except for the PCT ranking test: F4 sine T1, F4 complex T1 and T2 and F5 sine T1; the MCIm: 0.10 semitone at T2 and 0.05 semitone at T2. For this reason, the median and IQRs are presented. Of the 23 NHL, 19 were able to complete each test twice at intervals T1 and T2. The table and figures below present the DL scores for the PCTm and the SOECIC MTB PDT and percentage scores for the MCIm.

Table 7.4 NHL median and IQR scores for PCTm DL, SOECIC MTB PDT and MCIm
PCTm and SOECIC DLs in semitones, MCIm scores in %
n = 23 F4 = 349 Hz, F5 = 698 Hz

Test	Condition	T1			T2		
		Median	IQR	n	Median	IQR	n
PCTm discrimination	F4 sine	0.08	0.1	17	0.06	0.08	18
	F4 complex	0.07	0.1	22	0.05	0.05	18
	F4 piano	0.08	0.06	21	0.07	0.04	20
	F5 sine	0.08	0.07	21	0.07	0.06	18
	F5 complex	0.07	0.06	22	0.05	0.04	20
	F5 piano	0.06	0.04	20	0.06	0.05	20
PCTm ranking	F4 sine	0.04	0.1	14	0.05	0.08	14
	F4 complex	0.06	0.09	16	0.05	0.08	16
	F4 piano	0.07	0.04	18	0.06	0.08	16
	F5 sine	0.07	0.09	16	0.05	0.08	16
	F5 complex	0.06	0.03	19	0.05	0.06	18
	F5 piano	0.07	0.03	18	0.07	0.08	16
SOECIC MTB PDT	F4	0.16	0.32	23	0.21	0.09	19
MCIm	0.1	44.44	46.3	23	48.15	37.04	19
	0.05	25.93	48.16	23	33.33	42.6	19

7.3.7 Comparison of PCTm DL with SOECIC MTB PDT

Statistical comparisons were made between the SOECIC MTB PDT and PCTm discrimination scores as they shared the same task and unit. Scores were seen to be significantly affected by the test used to measure them (χ^2 (6) = 43.49, $p < .001$). Post hoc Wilcoxon analysis, with Bonferroni correction (0.05/21 = 0.0024), revealed significantly poorer scores for the SOECIC MTB PDT when compared with the PCTm discrimination scores, all 6 comparisons were $T = 0$, $p < 0.001$.

7.3.8 Musicianship

The sensitivity of each test to a known difference in ability, musicianship, was investigated. There were 14 musicians and 9 non-musicians within the NHL group.

Table 7.5 NHL musician and non-musician comparison using PCTm DL, SOECIC MTB PDT and MCIm Mann-Whitney U test. Musicians (n=14) non-musicians (n=9) T1 data only

		Non-musician median	Musician median	n	*U*	*p*	*r*
PCT discrimination (semitones)	F4 sine*	0.22	0.08	18	17.50	0.029	-0.32
	F4 complex*	0.15	0.06	22	18.50	0.004	-0.39
	F4 piano	0.15	0.08	21	28.50	0.067	-0.24
	F5 sine*	0.16	0.07	21	18.50	0.007	-0.37
	F5 complex*	0.15	0.06	22	12.50	0.001	-0.45
	F5 piano	0.08	0.07	20	37.50	0.367	-0.06
PCT ranking (semitones)	F4 sine	0.11	0.07	14	7.00	0.085	-0.28
	F4 complex*	0.09	0.07	16	5.00	0.029	-0.34
	F4 piano	0.10	0.07	18	14.00	0.079	-0.25
	F5 sine	0.09	0.08	16	12.00	0.182	-0.18
	F5 complex*	0.13	0.06	19	4.00	0.001	-0.47
	F5 piano	0.10	0.07	18	13.00	0.063	-0.27
SOECIC (semitones)	F4*	0.49	0.13	23	4.00	<0.001	-0.55
MCI (%)	0.1 semitone*	25.93	74.08	23	9.00	<0.001	-0.50
	0.05 semitone*	18.52	51.85	23	21.00	0.003	-0.39

*Significant at the $p < 0.05$ level, one-tailed

Although this allows comparisons between musicians and non-musicians per test, comparisons should not be made between the non-musician data for discrimination and ranking: more data was lost for the ranking than for the discrimination due to the DL calculation, because so many non-musicians had DLs that were calculated as over 0.25 semitone.

Both the SOECIC MTB PDT and the MCIm demonstrate sensitivity to musicianship for NHL, significantly poorer scores were seen with non-musicians compared to musicians. The PCT showed mixed sensitivity to musicianship for NHL. For the pitch discrimination tasks, the sine and complex tones at both F4 and F5 had significantly poorer scores for non-musicians compared to musicians. Interestingly, non-musicians performed surprisingly well for the piano stimuli for F4 and F5 (rather than musicians performing poorly), hence the lack of significant difference. For the pitch ranking tasks, only the complex stimuli showed a significant difference between musicians and non-musicians. The musicians performed at a similar level for both discrimination and ranking, and

although it seems that the non-musicians performed better in the ranking task when compared to the discrimination task, it is more likely that this apparent lack of difference is due to the greater data loss for non-musicians in the ranking task.

7.3.9 Reliability

This section presents the reliability of the PCTm using the DL scores, as well as reliability for the SOECIC MTB PDT and the MCIm, using the ICC (A,1). Use of the PCTm DL scores can be criticised due to the data loss associated with calculating the DL, and so a basic measure of agreement is also included in the section, which assesses the average differences between T1 and T2 psychometric curves.

Twenty-one of 23 NHL took part in each test twice, allowing test retest reliability of the PCTm DL scores to be assessed (1 NHL only managed to complete two subtests at T1 before withdrawing from the study, and 1 NHL did not return for any T2 testing). The numbers of pairs was reduced by the data loss associated with the DL. Test-retest reliability for the PCTm DL scores, SOECIC MTB PDT and MCIm was calculated using the ICC (A,1). As described in section 5.5, excellent reliability was defined as ≥ 0.8, and in addition, the critical value of *r* for *n* was used, e.g. the coefficient had to be higher than the amount of correlation that might be expected due to chance.

This meant that the reliability criteria for this experiment was determined by:

3. A coefficient of ≥ 0.8
4. A coefficient significantly greater than the critical value of *r* for *n*.

This criteria was met for 2/6 of the PCT discrimination conditions and for 3/6 of the PCT ranking conditions, 2/2 of the MCI conditions, and was not met for the SOECIC MTB PDT.

Table 7.6 NHL Intraclass correlation coefficient

n differed depending on comparisons. 95% confidence interval (ci) *ICC ≥ 0.8 and significantly greater than the critical value for *r*, for given *n*, at the p < 0.05 level (one-tailed)

Test	condition	n (pairs)	ICC(A,1)	95% ci	F (df1,df2)	sig
PCT discrimination	F4 sine*	18	0.804	0.414 - 0.935	3.603 (14,10)	0.022
	F4 complex	16	0.563	0.155 - 0.809	1.297 (17,18)	0.295
	F4 piano*	18	0.882	0.724 - 0.952	6.125 (18,19)	<0.001
	F5 sine	16	0.791	0.520 - 0.917	3.199 (17,17)	0.011
	F5 complex	18	0.599	0.226 - 0.819	1.530 (19,20)	0.178
	F5 piano	18	0.566	0.156 - 0.808	1.384 (18,18)	0.248
PCT ranking	F4 sine*	12	0.923	0.754 – 0.977	7.467 (11,11)	0.001
	F4 complex*	14	0.894	0.714 – 0.963	6.160 (14,15)	0.001
	F4 piano	14	0.649	0.255 – 0.860	1.578 (15,16)	0.188
	F5 sine	14	0.659	0.238 – 0.871	1.677 (14,13)	0.181
	F5 complex*	16	0.929	0.822 – 0.973	10.089 (17,18)	<0.001
	F5 piano	14	0.765	0.459 – 0.910	2.553 (16,17)	0.036
SOECIC MTB PDT		19	0.651	0.285 – 0.850	1.752 (18,18)	0.121
MCI (%)	0.1 semitone*	19	0.896	0.752 – 0.958	6.722 (18,19)	<0.001
	0.05 semitone*	19	0.917	0.801 – 0.967	8.620 (18,19)	<0.001

Due to the data loss from calculating the DL (see section 7.3.5), an alternative way to try and include as much data as possible was used in addition to the ICC (A,1). In order to compare psychometric curves, and include all usable data, each score per interval was compared between T1 and T2. The differences were obtained by subtracting T1 from T2 data, squaring it and square routing it, and then taking a mean for each stimulus type (see Appendix E for an example).

Table 7.7 Mean differences between T1 and T2 for NHL PCTm

PCTm discrimination			PCTm ranking		
Stimuli	Mean difference	n	Stimuli	Mean difference	n
F4 sine	2.20	19	F4 sine	2.44	20
F4 complex	3.10	21	F4 complex	2.31	21
F4 piano	1.74	21	F4 piano	2.13	21
F5 sine	1.91	19	F5 sine	2.40	20
F5 complex	1.80	21	F5 complex	1.75	21
F5 piano	2.08	21	F5 piano	2.41	21

7.3.10 Effect of stimulus type

The PCTm DLs for discrimination and ranking data were averaged over root note and stimulus type to investigate the effects of each using the Wilcoxon signed-rank test (*T*) and the Friedman's ANOVA (χ^2). To investigate the main effects of frequency (F4 and F5) and timbre (sine, complex and piano tones), the T1 PCT data was divided into F4 and F5 (timbres were combined) to investigate frequency, and divided into the 3 timbres (and frequencies were combined) to investigate timbre. This was done for pitch discrimination and pitch ranking separately. The Shapiro-Wilk test revealed that 9/10 variable distributions (two frequencies and three timbres for discrimination and ranking) differed significantly from normal for the pitch discrimination scores, and as such, non-parametric statistics were used.

The PCTm discrimination DL was significantly smaller for root note F5 (median = 0.07 semitone) compared to F4 (median = 0.08 semitone), $T = 573$, $p = 0.044$, $r = -0.18$. No significant effect of frequency was seen for the pitch ranking data ($p = 0.46$).

No significant effect of timbre was seen for the PCTm discrimination data set ($p = 0.063$). A significant effect of timbre was seen for the PCTm ranking data set, $\chi^2(2) = 16.15$, $p < 0.001$. Post hoc Wilcoxon, with Bonferroni correction ($0.05 \div 3 = 0.017$), and two tailed level of significance, revealed significantly better (lower) scores for the complex tone (Median (*Mdn)* = 0.07 semitones), when compared to the sine tone (*Mdn* = 0.08 semitones), $T = 59$, $p < 0.001$, r = -0.36 and the piano tone (*Mdn* = 0.08 semitones), T = 113, p = 0.002, r = -0.28.

7.4 CI Results

This section presents the results of the PCT test, and compares them with the results from the UW CAMP and the MCI tests, using a group of 22 CI users. Data from T2 was only used for the reliability analysis and the estimation of psychometric functions. As in section 7.3, the UW CAMP was analysed to determine whether the number of trials was sufficient to minimise the effects of chance on results. The effect of floor and ceiling effects were investigated for all 3 tests, psychometric functions were estimated and DLs were calculated and data loss was described. Results from the PCT and UW CAMP were compared. The effect of musicianship was analysed, test retest reliability was calculated and the effects of stimuli were investigated. Data from the UW CAMP was adjusted so that any scores of < 1 were reassigned the value of 1, except for when testing reliability on retest.

7.4.1 The role of chance in the UW CAMP

As described in section 7.3.1, adaptive methods result in differing numbers of trials per run. These trial numbers were considered in respect to the binomial distribution and the role of chance, and examples are presented for one good and one poor performer.

UW CAMP: 2 AFC, chance = 50%, 1 down 1 up, 3 repeats*

Good performer example

The following examples are those of a good performer, CI 1:

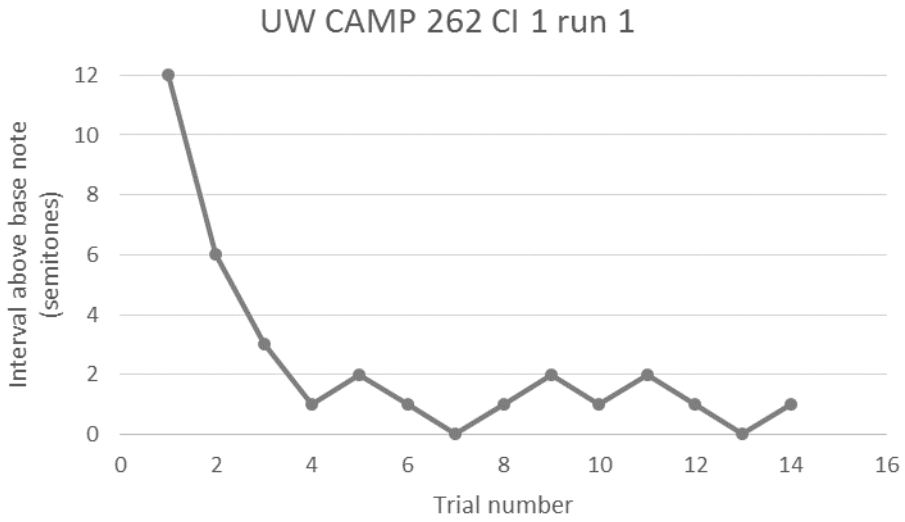


Figure 7.13 Example of a good performer (CI 1) using the UW CAMP Pitch Test (262 Hz), showing the adaptive staircase for run 1 only. Final score was 1 semitone.

Participant C1 was a good performer. The test's initial interval is always 12 semitones, and if successful it reduces to 6 semitones, and if successful it reduces further to 3 semitones. If 3 semitones is successful, the next interval is always 1 semitone. The adaptive procedure terminates after 8 reversals, and the score is the average of the final 6. The UW CAMP always includes 3 runs, and the final score for each base frequency is the average of the 3 runs.

In the example above, the final score was 1 semitone. The final 6 reversals (in reverse order) seen in Figure 7.13 were 1, 0, 2, 1, 2 and 0, and the average of these was 1 semitone.

Below is a breakdown of CI 1's performance, for each interval presented, the number of successes and the number of trials:

12 semitones: 1/1 successful

6 semitones: 1/1 successful

3 semitones: 1/1 successful

2 semitones: 3/3 successful

1 semitone: 3/6 successful

Using the binomial calculator, assuming a probability of success on a single trial being set at 0.5, the following numbers of successes are needed in order to keep the likelihood of success due to chance below 0.05:

Successes/trials	*p* of scoring that *due to chance*
Score = 5/5	0.03
Score = 6/6	0.016

This indicates that there were not enough successful trials presented for 12, 6, 3, 2 and 1 semitones in order for the alpha value of 0.05 to have been met. The final score was 1 semitone, and due to the low number of successful trials, the likelihood that this was due to chance is high.

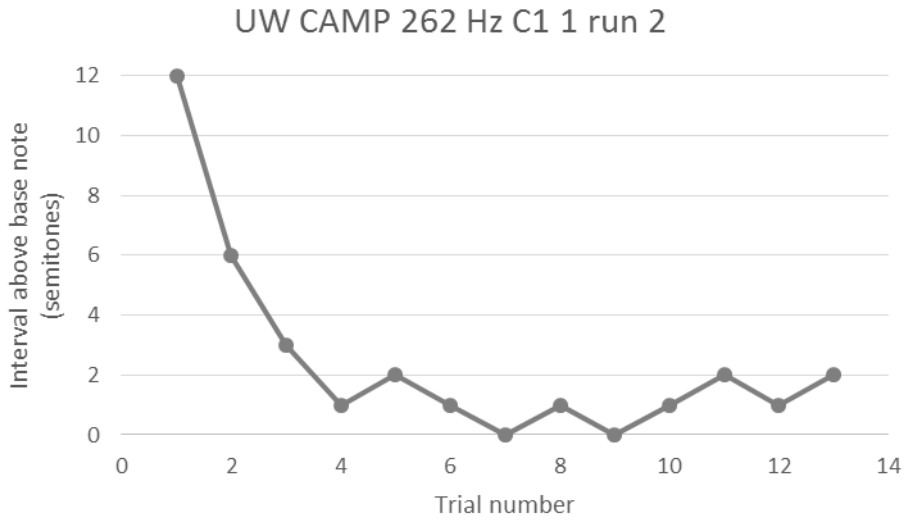


Figure 7.14 Example of a good performer (CI 1) using the UW CAMP Pitch Test (262 Hz), showing the adaptive staircase for run 2 only. Final score was 1 semitone.

In the example above, the final score was 1 semitone. The final 6 reversals (in reverse order) seen in Figure 7.14 were 2, 1, 2, 0, 1 and 0, and the average of these was 1 semitone.

12 semitones: 1/1 successful

6 semitones: 1/1 successful

3 semitones: 1/1 successful

2 semitones: 3/3 successful

1 semitone: 2/5 successful

This indicates that there were not enough successful trials presented for 12, 6, 3, 2 and 1 semitones in order for the alpha value of 0.05 to have been met. The final score was 1 semitone, and due to the low number of successful trials, the likelihood that this was due to chance is large.

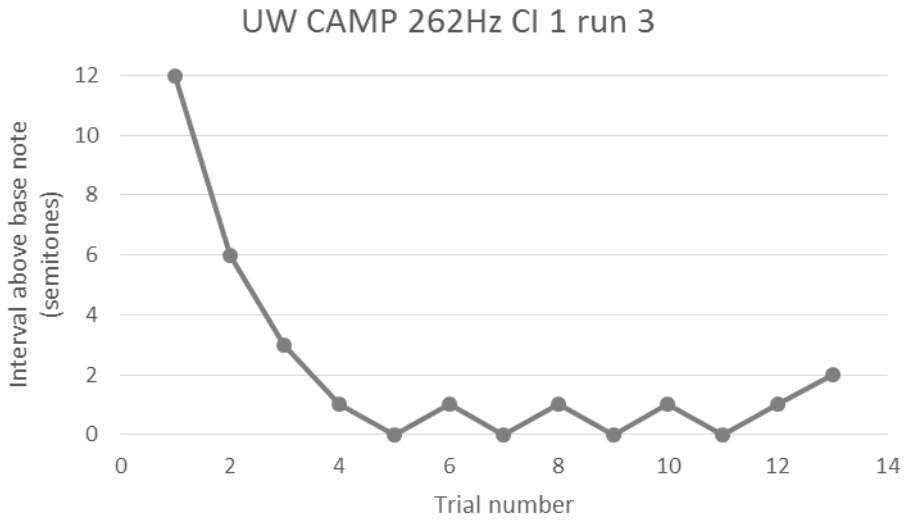


Figure 7.15 Example of a good performer (CI 1) using the UW CAMP Pitch Test (262 Hz), showing the adaptive staircase for run 3 only. Final score was 0.67 semitone (considered to be 1 semitone) due to the way in which the reversals are used to calculate a final average score.

In the example above, the final score was 0.67 semitone. The final 6 reversals (in reverse order) seen in Figure 7.15 were 2, 0, 1, 0, 1 and 0, and the average of these was 0.67 semitone.

12 semitones: 1/1 successful

6 semitones: 1/1 successful

3 semitones: 1/1 successful

2 semitones: 3/3 successful

1 semitone: 4/5 successful

This indicates that there were not enough successful trials presented for 12, 6, 3, 2 and 1 semitones in order for the alpha value of 0.05 to have been met. The final score was 1 semitone, and due to the low number of successful trials, the likelihood that this was due to chance is large.

When looking at each run individually, the number of trials presented is poor and the alpha value was often not met. When the trials are combined across runs, the numbers of trials are improved and are presented below:

Combined runs 1, 2 and 3:

12 semitones: 3/3 successful

6 semitones: 3/3 successful

3 semitones: 3/3 successful

2 semitones: 9/9 successful*

1 semitone: 9/16 successful

This indicates that there were only enough successful trials at 2 semitones in order for the alpha value of 0.05 to have been met. The final averaged score over the three runs was 0.89 semitone, which is considered to be representative of 1 semitone. Although this is not far from the interval 2 semitones, of which there were enough trials, there were not enough trials for the interval 1 semitone, in order for this score to be considered different from chance. Therefore, this would indicate that the algorithm to calculate the final score for the UW CAMP could be improved upon and the likelihood of this being due to chance is higher than $p < 0.05$ for the interval 1 semitone, casting doubt on this final score.

Poor performer example

The following examples are those of a poor performer, CI 21:

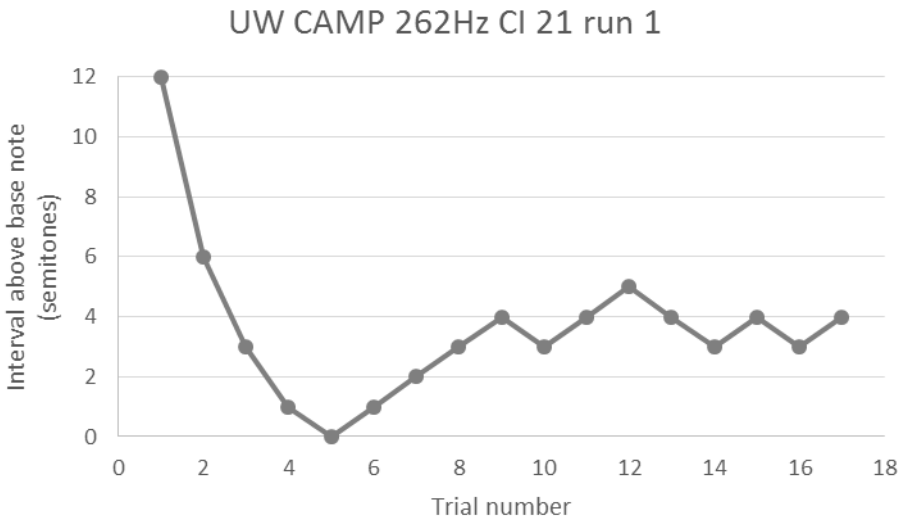


Figure 7.16: Example of a poor performer (CI 21) using the UW CAMP Pitch Test (262 Hz), showing the adaptive staircase for run 1 only. Final score was 3.67 semitones.

In the example above, the final score was 3.67 semitones. The final 6 reversals (in reverse order) seen in Figure 7.16 were 4, 3, 4, 3, 5 and 3 semitones, and the average of these was 3.67 semitones.

12 semitones: 1/1 successful

6 semitones: 1/1 successful

5 semitones: 1/1 successful

4 semitones: 4/5 successful

3 semitones: 1/5 successful

2 semitones: 0/1 successful

1 semitone: 1/2 successful

This indicates that there are not enough successful trials at any of the intervals tested in order for the alpha value of 0.05 to have been met. The final score was 3.67 semitones, and due to the low number of successful trials, the likelihood that this was due to chance is high.

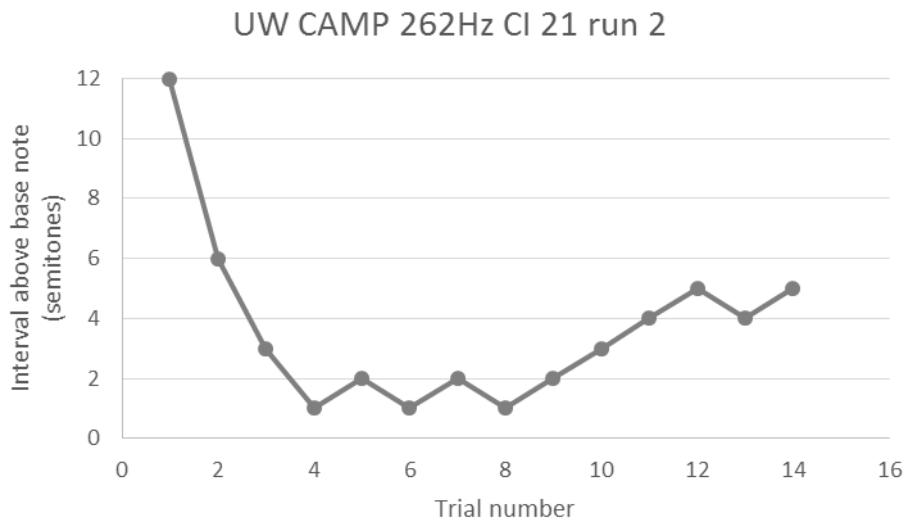


Figure 7.17 Example of a poor performer (CI 21) using the UW CAMP Pitch Test (262 Hz), showing the adaptive staircase for run 2 only. Final score was 3 semitones.

In the example above, the final score was 3 semitones. The final 6 reversals (in reverse order) seen in Figure 7.17 were 5, 4, 5, 1, 2 and 1 semitone, and the average of these was 3 semitones.

12 semitones: 1/1 successful

6 semitones: 1/1 successful

5 semitones: 1/2 successful

4 semitones: 0/2 successful

3 semitones: 1/2 successful

2 semitones: 2/3 successful

1 semitone: 0/3 successful

This indicates that there are not enough successful trials at any of the intervals tested in order for the alpha value of 0.05 to have been met. The final score was 3 semitones, and due to the low number of successful trials, the likelihood that this was due to chance is high.

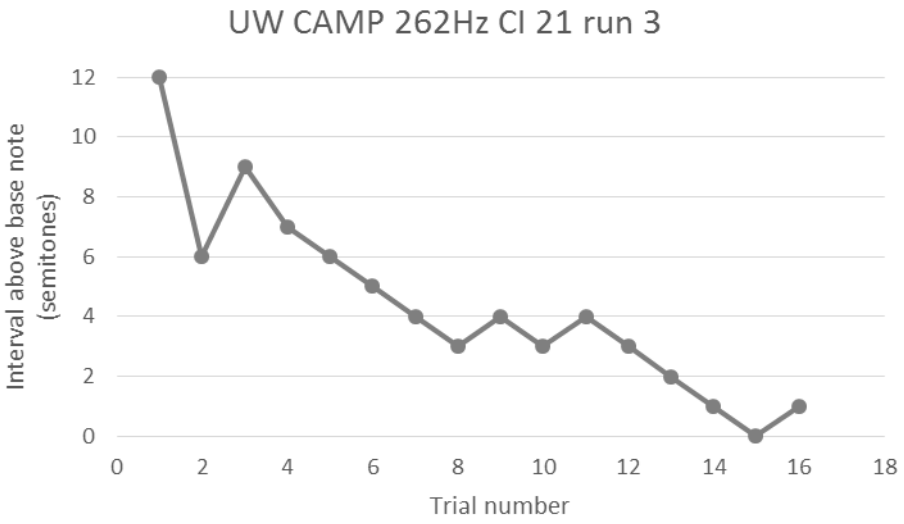


Figure 7.18 Example of a poor performer (CI 21) using the UW CAMP Pitch Test (262 Hz), showing the adaptive staircase for run 3 only. Final score was 2.5 semitones.

In the example above, the final score was 2.5 semitones. The final 6 reversals (in reverse order) seen in Figure 7.18 were 1, 0, 4, 3, 4 and 3 semitones, and the average of these was 2.5 semitones.

12 semitones: 1/1 successful

9 semitones: 1/1 successful

7 semitones: 1/1 successful

6 semitones: 1/2 successful

5 semitones: 1/1 successful

4 semitones: 3/3 successful

3 semitones: 1/3 successful

2 semitones: 1/1 successful

1 semitone: 1/2 successful

This indicates that there are not enough successful trials at any of the intervals tested in order for the alpha value of 0.05 to have been met. The final score was 2.5 semitones, and due to the low number of successful trials, the likelihood that this was due to chance is high.

Combined runs 1, 2 and 3:

12 semitones: 3/3 successful

9 semitones: 1/1 successful

7 semitones: 1/1 successful

6 semitones: 3/4 successful

5 semitones: 3/4 successful

4 semitones: 7/10 successful

3 semitones: 3/10 successful

2 semitones: 3/5 successful

1 semitone: 2/7 successful

This indicates that even when runs 1, 2 and 3 were combined, there were not enough successful trials at any intervals in order for the alpha value of 0.05 to have been met. The final score was 3.06 semitones and due to the low number of successful trials, the likelihood that this was due to chance is high.

7.4.2 Floor and ceiling effects

As in Chapter 5, ceiling effects were defined as the best possible score that is achievable per test, and floor effects were defined as the lowest possible score (for adaptive methods) or scores at chance level per test. For the UW CAMP, the maximum score was 12 semitones, and the lowest score was 1 semitone, with ceiling effects being defined as any score above 11 semitones for this test because there wasn't much accuracy between 11 and 12 semitones. The maximum score for the MCI was 100% and chance level was 11%. Floor and ceiling effects were analysed using the raw data from the PCT and as such the definition was adapted in order to encompass data from all intervals. Rather than define ceiling as 100% for all intervals, ≥80% was used instead, as this would otherwise seemingly give an unfair advantage to the PCT, plus the ceiling effects that were seen using this definition typically had a score of ≥80% for the smallest interval and 100% for all other intervals. Floor effects were defined as ~≤ chance level, which was 50% for the PCT discrimination and 25% for PCT ranking. These were reported for an individual if it occurred for at least 1/6 conditions. Floor and ceiling effects were reported for T1 data only, as the T2 data was only gathered in order to perform reliability analyses, and in a clinical setting, pitch tests would typically only be performed once due to time constraints.

PCT discrimination

Ceiling effects (≥80% for all intervals) were seen for 5/22 CI users: CI 1 (F4 piano, F5 complex), CI 5 (F5 complex), CI 13 (F4 complex, F5 sine, complex, piano), CI 17 (F5 complex) and CI 20 (F5 sine). Floor effects (~≤50% for all intervals) were seen for 1/22 CI user: CI 8 (F4 sine).

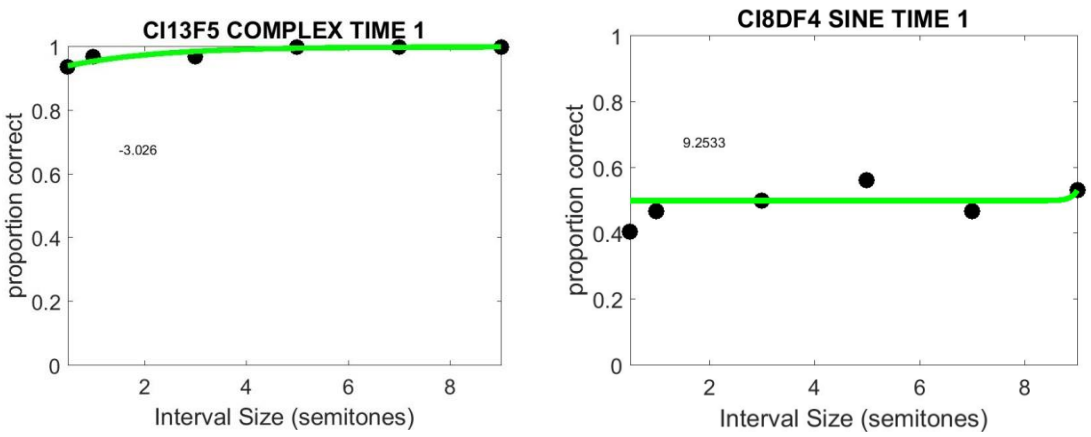


Figure 7.19 Examples of floor and ceiling effects. Ceiling: PCT ranking, CI 13, F5 complex, ≥80%. Floor: PCT discrimination, CI 8, F4 sine, ~≤50%.

PCT ranking

Ceiling effects (≥80% for all intervals) were seen for 3/22 CI users: CI 1 (F5 complex), CI 13 (F5 sine, complex) and CI 17 (F5 complex). Floor effects (~≤25% for all intervals) were seen for 3/22 CI users: CI 16 (F4 piano), CI 19 (F4 piano) and CI 21 (F5 sine).

UW CAMP

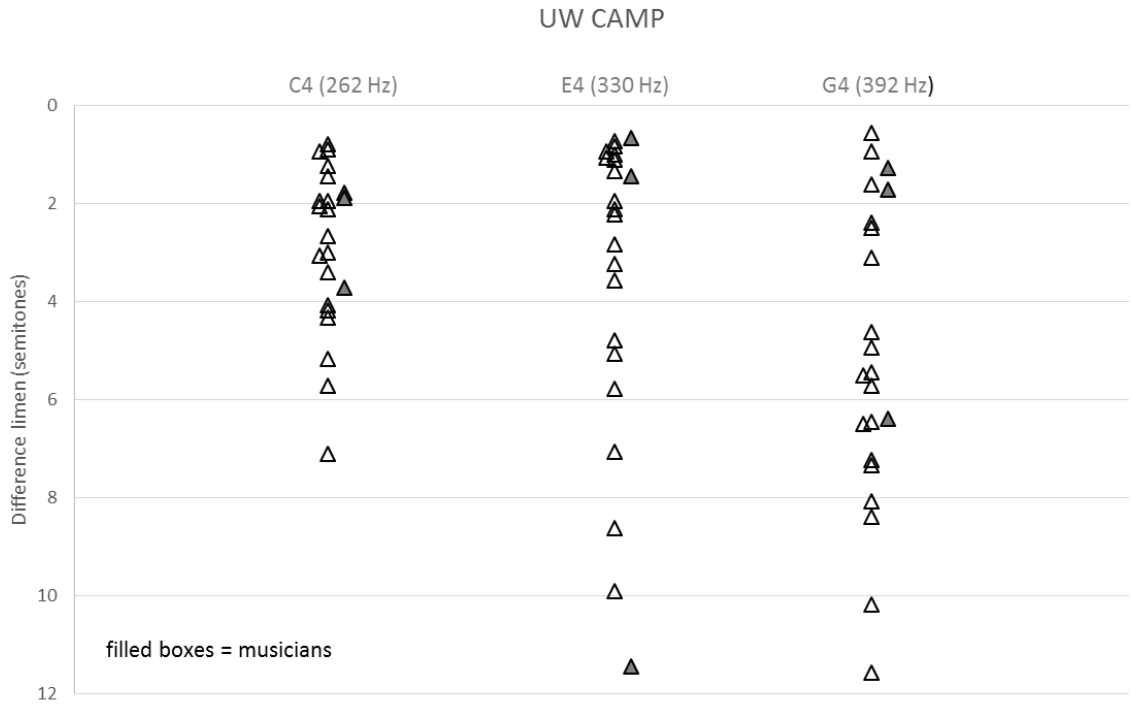


Figure 7.20 CI user UW CAMP scores showing range of scores and the effect of musicianhsip

UW CAMP 262 Hz: ceiling effects (≤1 semitone) were seen for 3/22: CI 1, 9, 11. No floor effects (≥11 semitones) were seen.

UW CAMP330 Hz: ceiling effects (≤1 semitone) were seen for 5/22 CI users: CI 1, 4, 10, 14, 15. Floor effects (≥11 semitones) were seen for 1/22 CI user: CI 13.

UW CAMP 392 Hz: ceiling effects (≤1 semitone) were seen for 2/22 CI users: CI 1, 4. Floor effects (≥11 semitones) were seen for 1/22 CI user: CI 5.

MCI

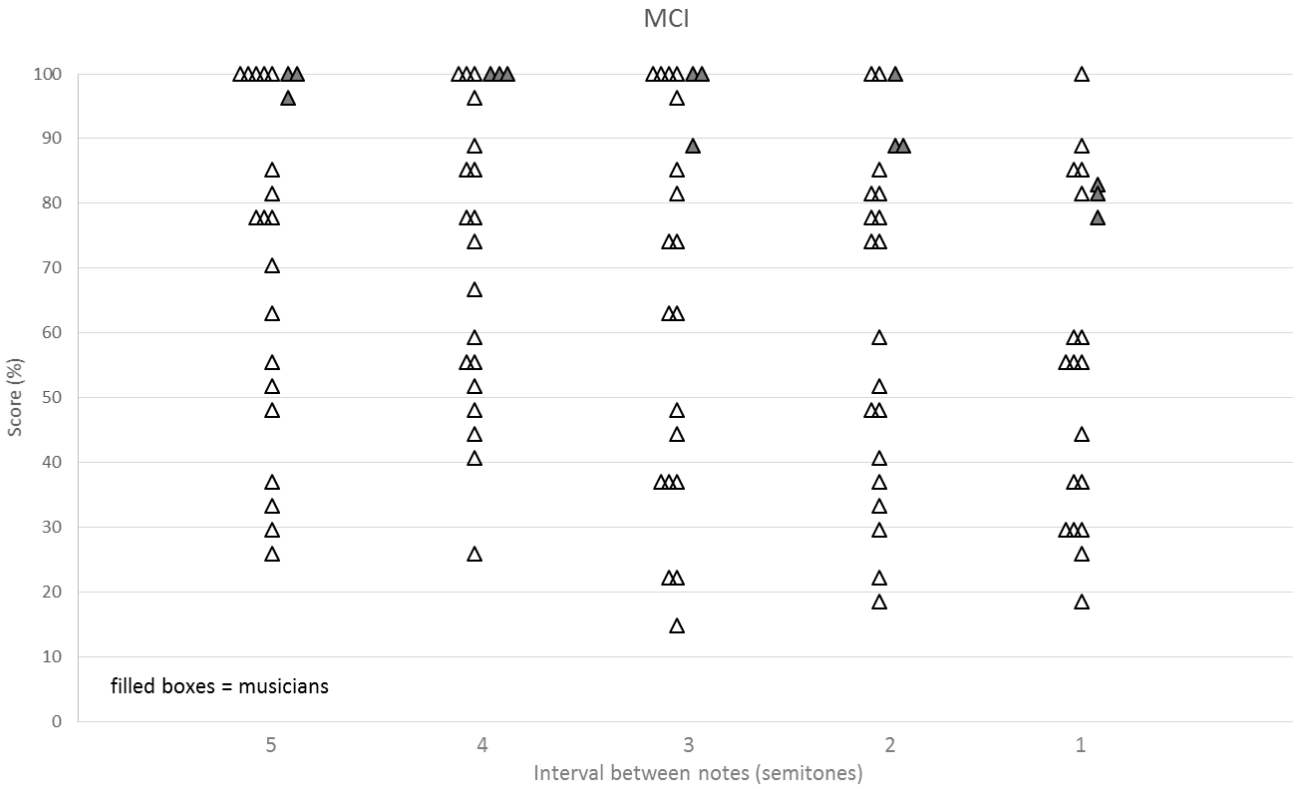


Figure 7.21 CI user MCI scores for the intervals 1 – 5 semitones showing a range of scores and effect of musicianship

MCI 5: Ceiling effects (100%) were seen for 7/22 CI users: CI 1, 5, 9, 11, 14, 15, 16. No floor effects (≤11%) were seen.

MCI 4: Ceiling effects (100%) were seen for 6/22 CI users: CI 1, 5, 12, 13, 15, 17. No floor effects (≤11%) were seen.

MCI 3: Ceiling effects (100%) were seen for 6/22 CI users: CI 1, 5, 12, 13, 16, 17. No floor effects (≤11%) were seen.

MCI 2: Ceiling effects (100%) were seen for 3/22 CI users: CI 1, 5, 17. No floor effects (≤11%) were seen.

MCI 1: Ceiling effects (100%) were seen for 1/22 CI user: CI 1. No floor effects (≤11%) were seen.

7.4.3 Psychometric functions for CI

Psychometric functions were fitted to the data as described in section 7.3.3. Seventeen of the 22 participants attempted all 6 conditions at T1 and T2, and the remaining 5/22 were only able to attempt 3 conditions at T1 and T2. It was realised after testing CI 6 at F4, and CI 4 with F5, that the sine tone stimuli that had been used until that point had been accessed from the incorrect files, and some loudness cues may have been audible between the two different notes within the triplet, in the region of 5dB(A). It was decided to discard the data for CI 1 – 6, for F4 sine, and CI 1 – 4, for F5 sine. As such, 214 psychometric functions were generated for the PCT discrimination test and 214 for the ranking test, making a total of 428 psychometric functions. The majority of these were monotonic; better performance for larger intervals and poorer performance for smaller intervals. Examples of monotonic psychometric functions can be seen below in Figure 7.22.

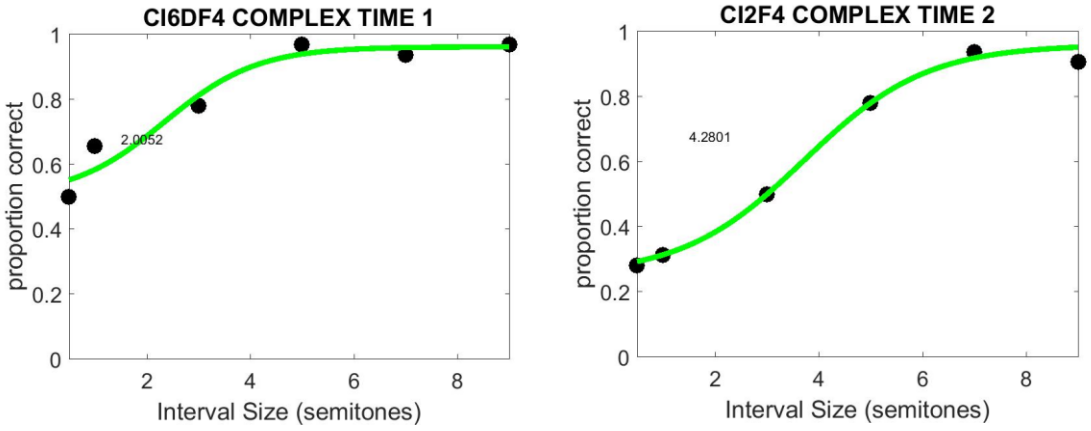


Figure 7.22 Two examples of CI user monotonic psychometric functions

Non-monotonic functions were defined using the following subjective criteria: scores that did not follow either a positive or a negative trend across the intervals, e.g. often showing better scores at smaller rather than larger intervals. Only one example of a truly 'reversed' function was seen (Appendix C: CI 17 F4 piano T2), with best performance for the smallest intervals and poorest performance at the largest intervals. Flat scores across intervals were not counted as non-monotonic. Functions were labelled as non-monotonic individually; e.g. a non-monotonic label did not require both T1 and T2 to display non-monotonic features. There were 35 non-monotonic

functions seen in the CI group (9 for discrimination and 26 for ranking), which is 8.2% (35/428). All 35 non-monotonic psychometric functions for CI users can be seen in Appendix C.

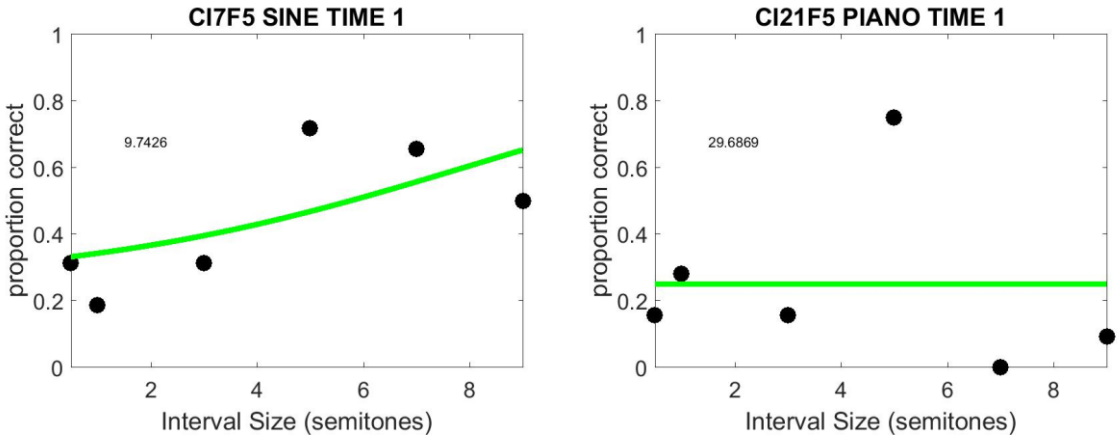


Figure 7.23 Examples of PCT ranking non-monotonic psychometric functions with CI users

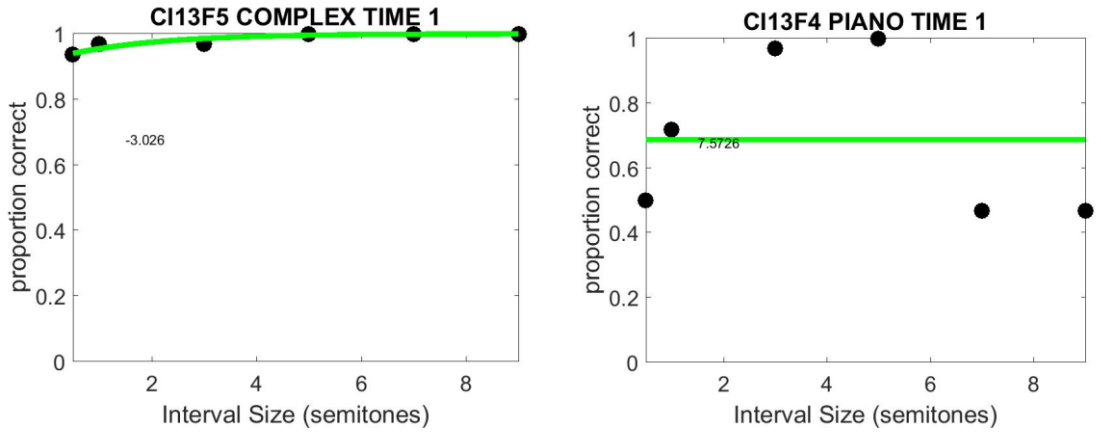


Figure 7.24 Examples of two psychometric functions from the PCT pitch ranking test, for CI 13 (a musician) that were rejected: the function on the left shows such good performance that the difference limen was calculated as being negative, and the function on the right showed a non-monotonic function and the difference limen was not in keeping with the spread of the data.

Of the 9 non-monotonic discrimination functions, 4 were seen in the sine stimuli, 2 were seen in the complex stimuli and 3 were seen in the piano stimuli. Of the 26 non-monotonic ranking functions, 3 were seen in the sine stimuli, 2 were seen in the complex stimuli and 21 were seen in the piano stimuli. Non-monotonic psychometric functions occurred for 12 of the 22 CI users,

confirming that adaptive staircase procedures would be inappropriate and could lead to erroneous values for 55% of this sample of CI users.

Very low scores of close to 0 correct for either discrimination or ranking indicate that the participant is scoring much more poorly than chance, indicating that they are actively pursuing the opposite of the correct response. For the discrimination task, this might indicate a lapse in concentration, however for the ranking task, it indicates that the pitches are being reversed. Of the 428 psychometric functions, 13 reversals were seen, 12 of which were seen in the ranking task, indicative of pitch reversals (see Appendix D) and were mostly from the piano stimuli. Six of the 22 CI users showed these pitch reversals for at least one condition.

7.4.4 PCT difference limen

As described above in section 7.3.4, the PCT was specifically designed with the use of the MCS in order to make no assumptions about CI users' functions for pitch perception. This means that an individual may score better for some intervals than for others and the nature of this may not be predictable, e.g. it may not be monotonic. In order to provide feedback to participants plus allow comparisons with other tests, a cut off score of 22/32 was used to indicate that that interval was 'successfully' discriminated or ranked (as described in Chapter 6). This cut off point applies to each interval independently, and one of the strengths of the PCT is that it provides this information for each interval it tests, and does not make the assumption that if a participant can successfully pitch rank at 5 semitones, then they will be able to do it at 7 or 9 semitones too.

In order to compare with other tests, the DL was calculated using the MATLAB Palamedes Toolbox (Prins and Kingdom, 2009), as described in section 7.3.4. These DL scores were then plotted so that the range of DL data could be visualised (Figure 7.25).

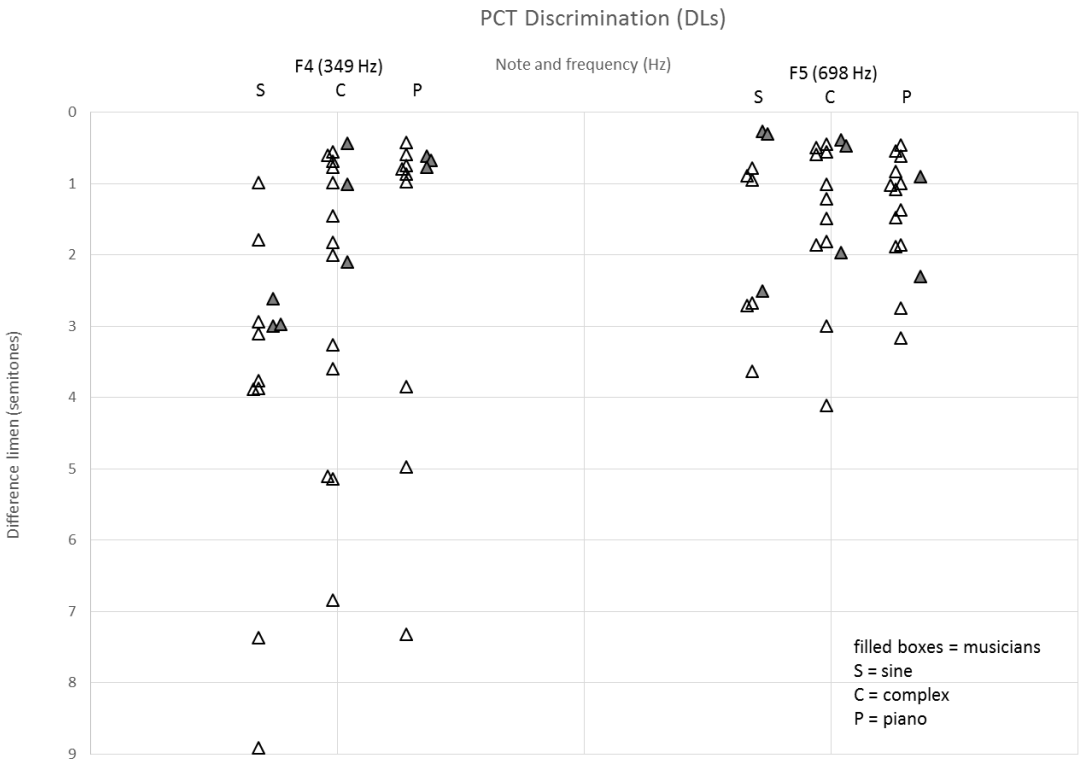


Figure 7.25 CI user PCT discrimination DL results showing the range of scores and the effect of musicianship. Note: musicians n = 3, however some data was lost due to the DL.

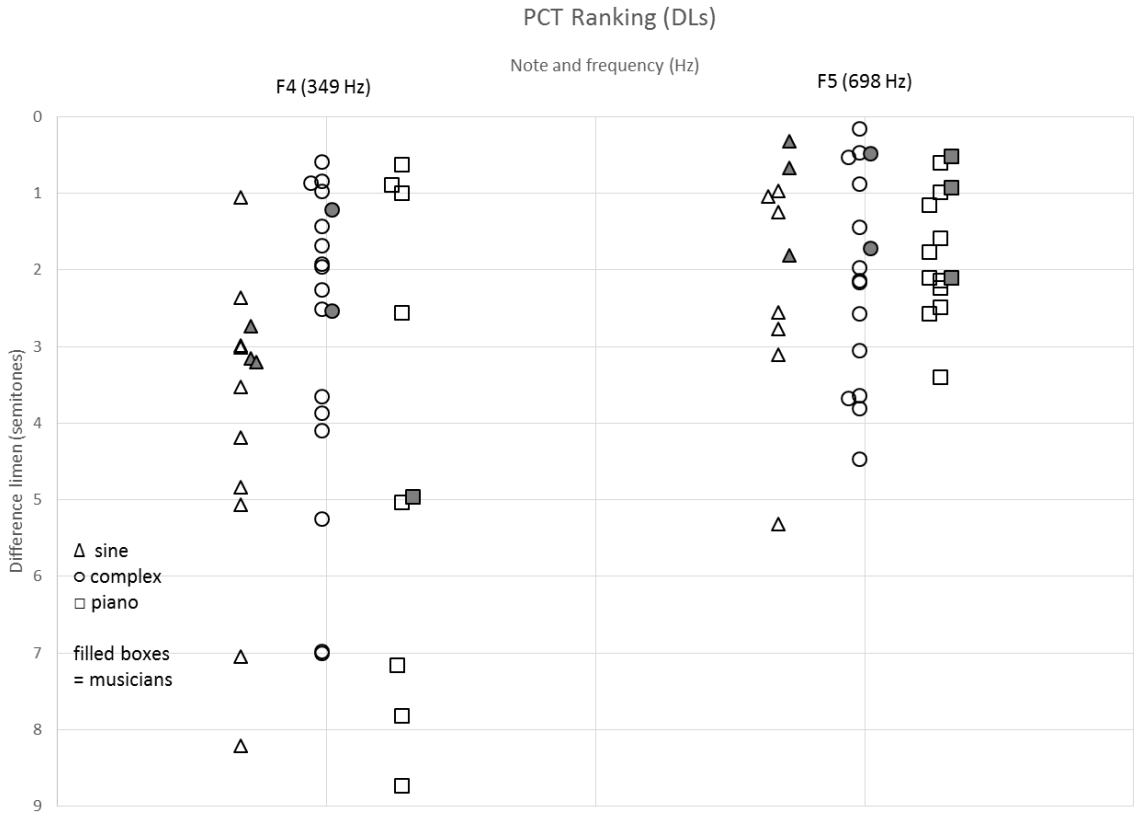


Figure 7.26 CI user PCT ranking results showing the range of scores and the effect of musicianship. Note: musicians n = 3, however some data was lost due to the DL.

7.4.5 Data loss due to the DL

As described in the NHL section above, calculating a threshold in this way highlighted further complexities in the data. In 17 cases (3.6%) the DL was negative, and so these were rejected. The bounds of the PCT were 0.5 and 9 semitones, and so 9 was used as an upper bound, which meant 47 (10%) were rejected (12 from discrimination and 35 from ranking). Each DL below 0.5 semitones was assessed individually for visual goodness of fit, which resulted in 24 (5.1%) being rejected (16 from discrimination and 8 from ranking). The total loss of CI user data due to the DL calculation was 18.7%.

7.4.6 Median scores for the PCT, UW CAMP and the MCI

The Shapiro-Wilk test of normality was used due to its ability to cope with low numbers of *n* (Shapiro and Wilk, 1965). The majority of the difference limens were significantly non-normally distributed and so a non-parametric statistical approach was taken. For this reason, the median and IQRs are presented. All 22 CI users were able to complete each test twice at intervals T1 and T2. The table and figures below present the DL scores for the PCT and the UW CAMP and percentage scores for the MCI.

Table 7.8 CI user median DL scores with interquartile range (IQR)
for the PCT, CAMP and MCI tests (PCT and CAMP DLs in semitones, MCI scores in %).
N = 22. Stimulus type: F4 = 349 Hz, F5 = 698 Hz, s = sine, c = complex and p = piano tone.

Test	Condition (note, stimulus type or interval)	T1 Median	IQR	T2 Median	IQR
PCT discrimination	F4 sine	3.05	11.18	3.07	2.22
	F4 complex	1.45	2.49	1.60	2.28
	F4 piano	0.83	3.31	1.30	1.93
	F5 sine	1.16	1.78	0.97	1.28
	F5 complex	1.11	1.33	1.26	1.20
	F5 piano	1.09	1.07	1.41	1.32
PCT ranking	F4 sine	3.20	1.69	4.03	1.60
	F4 complex	2.25	2.60	2.88	3.77
	F4 piano	5.00	6.17	3.88	3.20
	F5 sine	1.53	1.67	1.70	1.61
	F5 complex	2.13	2.56	2.07	2.48
	F5 piano	2.10	1.21	1.57	1.93
CAMP	262 Hz	2.39	32.17	1.70	42.81
	330 Hz	2.17	43.19	1.61	43.12
	392 Hz	5.47	74.62	5.45	85.27
MCI	5	77.78	47.22	85.19	48.15
	4	77.78	43.52	85.19	40.74
	3	74.07	60.19	92.59	44.44
	2	74.07	41.67	77.78	48.15
	1	55.56	44.44	68.52	49.08

7.4.7 Comparison of PCT DL with UW CAMP

Statistical comparisons were made between the UW CAMP and PCT ranking scores as they shared the same task and unit. Scores were seen to be significantly affected by the test used to measure them (χ^2 (8) = 18.4, p = .018). Post hoc Wilcoxon analysis, with Bonferroni correction (0.05/36 = 0.0013), revealed significantly poorer scores for the UW CAMP 392 Hz (G4) when compared to the UW CAMP 262 Hz (C4), (T = 35, p = 0.002, r = -0.45), the PCT F4 complex (T = 10, p < 0.001, r = -

0.55), the PCT F5 complex ($T = 0$, $p < 0.001$, $r = -0.62$) and the PCT F5 piano ($T = 6$, $p = 0.002$, $r = -0.55$).

7.4.8 Musicianship

Figure 7.20, Figure 7.21, Figure 7.25 and Figure 7.26 show the performance of the three musicians (defined as holding a musical qualification and/or deeming yourself a musician) within the group. Three musicians within a group of 22 was considered too small to statistically analyse however the trends are interesting. Musicians were top performers in the MCI, and performed well within both parts of the PCT. Interestingly, the musician results within the UW CAMP are scattered, which may be caused by the adaptive procedure.

7.4.9 Reliability

All 22 CI users took part in each test twice. Test-retest reliability for the PCT DL scores, UW CAMP and MCI was calculated using the ICC (A,1). As described in section 5.5, excellent reliability was defined as ≥ 0.8, and in addition, the critical value of *r* for *n* was used, e.g. the coefficient had to be higher than the amount of correlation that might be expected due to chance.

This meant that the reliability criteria for this experiment was determined by:

5. A coefficient of ≥ 0.8
6. A coefficient significantly greater than the critical value of *r* for *n*.

This criteria was met for 4/6 of the PCT discrimination conditions and for 4/6 of the PCT ranking conditions, 5/5 of the MCI conditions, and only 1/3 of the CAMP conditions, G4 (392 Hz).

Table 7.9 Test-retest reliability (ICC) analyses for the PCT, CAMP and MCI

N = 22. * indicates ICC > 0.8 and significantly different from the critical value of r for n. Stimulus type: s = sine, c = complex and p = piano tone. ICC(A,1), two way random, absolute agreement, single measures, F test 2 tailed 0.05

Test	Condition	n (pairs)	ICC (A,1)	95% ci	F(df1, df2)	sig
PCT	F4 sine*	14	.818	.501 – .941	3.30 (12,12)	.034
discrimination	F4 complex	17	.735	.407 – .895	2.27 (16,16)	.054
	F4 piano*	13	.941	.797 – .984	9.29 (10,10)	.001
	F5 sine*	10	.901	.639 – .975	4.45 (9,9)	.019
	F5 complex*	13	.855	.572 – .956	3.618 (11,11)	.021
	F5 piano	16	.434	-.109 – .770	0.85 (14,14)	.614
PCT ranking	F4 sine*	13	.858	.594 – .955	3.70 (12,12)	.016
	F4 complex*	18	.816	.579 – .927	3.56 (17,18)	.006
	F4 piano	9	.776	.197 – .951	1.58 (7,7)	.280
	F5 sine*	10	.984	.940 – .996	27.64 (9,10)	<.001
	F5 complex*	16	.818	.539 – .935	3.31 (14,14)	.016
	F5 piano	14	.796	.481 – .929	2.67 (13,13)	.042
CAMP	262 Hz	22	.585	.226 – .804	1.58 (21,21)	.149
	330 Hz	22	.453	.072 – .725	1.11 (21,22)	.408
	392 Hz*	22	.851	.668 – .936	5.34 (21,21)	<.001
MCI	5*	21	.931	.836 – .972	11.41 (20,21)	<.001
	4*	21	.900	.770 – .952	7.40 (20,20)	<.001
	3*	21	.868	.583 – .952	6.22 (20,12)	<.001
	2*	21	.928	.831 – .970	10.37 (20,20)	<.001
	1*	20	.854	.667 – .940	4.82 (19,19)	<.001

Due to the data loss from calculating the DL (see 7.3.5), an alternative way to try and include as much data as possible was used in addition to the ICC (A,1). In order to compare psychometric curves, and include all usable data, each score per interval was compared between T1 and T2. The differences were obtained by subtracting T1 from T2 data, squaring it and square routing it, and then taking a mean for each stimulus type (see Appendix E). The results are presented in Table 5.10 below.

Table 7.10 Mean differences between T1 and T2 for CI PCT

PCTm discrimination			PCTm ranking		
Stimuli	Mean difference	n	Stimuli	Mean difference	n
F4 sine	1.67	16	F4 sine	2.63	22
F4 complex	2.00	22	F4 complex	2.80	22
F4 piano	2.41	22	F4 piano	3.99	22
F5 sine	0.77	13	F5 sine	1.21	17
F5 complex	1.20	17	F5 complex	1.72	17
F5 piano	1.30	17	F5 piano	1.53	17

7.4.10 Effect of stimulus type

The Shapiro-Wilk test revealed that 8/10 variable distributions (two frequencies and three timbres for discrimination and ranking) differed significantly from normal for the pitch discrimination scores, and as such, non-parametric statistics were used. The DLs for discrimination and ranking data were averaged over root note and stimulus type to investigate the effects of each using the Wilcoxon signed-rank test (*T*) and the Friedman's ANOVA (χ^2). To investigate the main effects of frequency (F4 and F5) and timbre (sine, complex and piano tones), the T1 PCT data was divided into F4 and F5 (timbres were combined) to investigate frequency, and divided into the 3 timbres (and frequencies were combined) to investigate timbre. This was done for pitch discrimination and pitch ranking separately.

The PCT discrimination DL was significantly smaller for root note F5 (median = 1.03 semitones) compared to F4 (median = 2.01 semitones), $T = 130$, $p = .02$, $r = -0.29$. The PCT ranking DL was significantly smaller for root note F5 (median = 1.89 semitones) compared to F4 (median = 3 semitones), $T = 89$, $p = .008$, $r = -0.35$.

An effect of stimulus type was seen for the PCT discrimination DLs ($\chi^2(2) = 7.41$, $p = .025$). Post hoc Wilcoxon signed rank test with Bonferroni correction (*sig* of .0167) revealed that the only significant difference was seen between the sine (median = 2.69 semitones) and the piano tones (median = 0.99 semitones), $T = 28$, $p = .005$, $r = -0.37$. No effect of stimulus type was seen for the PCT ranking DLs.

A root note effect was also observed for the UW CAMP ($\chi^2(2) = 6.23$, $p = .047$); however, contrary to the PCT, larger DLs were observed for the highest root note, 392 Hz (median = 5.47 semitones) when compared to the lowest note, 262 Hz (median = 2.39 semitones) ($T = 35$, $p = -.002$, $r = -0.45$).

7.5 Discussion

The overall aim of this chapter was to assess whether the PCT can be considered to be a suitable test for CI users, whether it showed improvement on existing tests, and what information it can provide to advance knowledge about the pitch perception capabilities of CI users. This experiment has comprised two parts; a study using NHL and a study using CI users. The use of NHL was beneficial in two ways. It enabled the PCT to be tested prior to use with CI users, and the inclusion of microtonal stimuli within the PCTm (and within the MCIm) meant that the NHL data could be seen as an approximate model of the CI data.

It was hypothesized that the PCT would perform as well, if not better than the MCI, UW CAMP and SOECIC MTB PDT, due to the fact that the PCT was designed based on the best features of existing tests:

1. Use of the MCS
2. A large number of repeats in order to provide statistical confidence
3. A suitable level of difficulty in order to avoid floor and ceiling effects
4. A simplified number and complexity of contours, when compared to the MCI

In addition the PCT provides a method to assess pitch discrimination and pitch ranking ability simultaneously, which is unique to the PCT. The research questions will be addressed in order here, and issues and limitations will be discussed.

7.5.1 Does the PCT provide enough repeats to give statistical confidence in the final result?

As introduced in Chapter 4, for a test to provide statistical confidence in the results, it must firstly have enough trials, and secondly, have a predefined level of success in order that a result can be said to be a success at a certain level of difficulty. By design, the PCT provided this. For each interval, a participant needed to achieve 22/32 or greater in order for the interval to be considered to be successful. This cut off score of 22 or greater meant that the possibility of that score or better occurring by chance was less than 5%. The PCT therefore provided enough repeats per condition, and in addition specified the required level for success. This meant that test users could clearly interpret results when providing feedback for test participants.

The 32 trials per interval per condition was made up of 8 repeats for each of the 4 contours. The original reason for having 8 repeats per contour was that this allowed one mistake to be made (e.g. a lapse in concentration) and even with a score of 7/8, the likelihood of this being due to chance alone was below 5%. Due to the possibility that intervals may be perceived differently depending on the contour shape, it was decided that each contour should have 8 repeats, thus meaning that for any given interval, there were 32 repeats. A recalculation using a binomial calculator showed that a score of 22 or better out of 32 was high enough to consider that interval successfully perceived, whilst keeping the likelihood of this being due to chance, below 5%.

These calculations were made with an assumption of chance being 50%, which is the case for the discrimination task, rather than 25%, which is the case for a ranking task. The experimenter originally thought that only one level of chance could be used, and this is why the 50% level was used, however more recently it has been realised that the discrimination and ranking chance level scores could be scored independently. This meant that the PCT ranking scores would only need to achieve a score of 13 or more out of 22 in order to keep the probability below 0.05, and this experiment has set this level at 22/32. This means that the PCT ranking scores would actually be better than the results reported here, if they were to be rescored, and indicates that the results presented here are much more conservative than they may have been, and that thresholds for pitch ranking scores would actually be lower indicating better performance.

As discussed in section 5.7.1, generally the pitch perception tests have shown enough trials for both MCS and adaptive procedures when used with NHL and CI users, except for some examples (e.g. MedEl did not show enough trials for good performing NHL, UW CAMP not for poor performing CI users). Evidence from the current experiment shows that whilst the SOECIC MTB PDT (with NHL) has generally shown enough repeats to keep the likelihood of the results being due to chance below 5% in the region of asymptote (see section 7.3.1), the UW CAMP (with CI users) has not (section 7.4.1). Detailed analysis was not carried out for every presentation and for every participant, however a typical ‘good’ and ‘poor’ performer were chosen from each test in order to represent the participant group.

The SOECIC MTB PDT terminates after 7 reversals and calculates the final score by taking an average of the last 5. Both good and poor performers were presented with plenty of repeats around the area of calculated threshold: good performer NHL 8 was successful 5/7 times at 16 cents and the calculated final score was 18 cents; poor performer NHL 14 was successful 4/5 times at 100 cents and the calculated final score was 106 cents. Similar high numbers of repeats were seen with the SOECIC MTB PDT with CI users in Chapter 5.

In comparison, the UW CAMP terminates after 8 reversals, and calculates the final score by taking an average of the last 6. Both the good and poor performer illustrations in section 7.4.1 do not achieve enough successful trials around the calculated threshold, even when the 3 runs are taken into account. There are a number of reasons as to why this was achieved successfully for the SOECIC MTB PDT and not the UW CAMP: the UW CAMP is 2 AFC rather than 3, the level of chance is 50%, meaning that any binomial calculations are going to be more stringent; the UW CAMP uses a 1 down 1 up staircase, meaning that inevitably there will be less presentations than the SOECIC MTB PDT which uses a 2 down 1 up procedure; the SOECIC MTB PDT is testing pitch discrimination whereas the UW CAMP is testing pitch ranking; and finally, the SOECIC MTB PDT is able to present intervals as small as 1 cent, and as this is well below the capabilities of both CI users and NHL, there is no need to add any kind of 'reversal' at '0' semitones.

Finally, the MCI does not report on a recommended number of required trials to be confident in the results. Galvin *et al. (*2007) report that they repeated the original version of the MCI (e.g. 9 contours, 5 semitone intervals, 3 root notes = 135 unique trials) until their participants reached stability (Galvin *et al.,* 2007, p 307), and then the average score was calculated. For Galvin *et al's* subject 'S1', this was only 2 repeats, but for their subject 'S4', this was 16 repeats. As the MCI uses a 9 AFC, (theoretical) chance levels = 11%, which means that the likelihood of achieving any scores due to chance alone are much more reduced than if the test was a 2 or 3 AFC test.

Overall, the PCT provides a large number of trials in order to keep the effects of chance to a minimum, whilst also allowing for lapses in concentration, and specifies the number required for 'success' at a given level of difficulty. This gives it an advantage over the UW CAMP which due to its design is more likely to be affected by chance. The PCT is similar in nature to the MCI, however the way that the PCT is presented means that the number of repeated trials which are deemed necessary are always presented to every participant, which is unlike the MCI.

7.5.2 Does the PCT provide a suitable level of difficulty for test users?

Suitable levels of difficulty in pitch tests for CI users are important because inappropriate difficulty means that ability range cannot be defined and changes in ability (improvements or deteriorations) cannot be assessed successfully. For the PCT, ceiling effects were defined as all intervals being above 80%, rather than the maximum score possible; this was because of the difficulty range of the intervals, it was unlikely that participants would be able to score 100% at the smallest intervals presented, but that scores very near to 100% performance would still mean that the test would be poor at measuring any improvement. It wasn't so easy to apply these criteria to the UW CAMP, MCI or the SOECIC MTB PDT as there wasn't an obvious way to make a similar comparison, hence

for all tests other than the PCT, maximum and minimum possible scores, or chance scores were used to define ceiling and floor effects. Although not strictly a fair comparison, this means that if any test is disadvantaged by this approach, it would be the PCT, rather than the other tests.

The PCT was designed specifically to present pitch intervals that were within a suitable range of difficulty for CI users. Guidance from the literature facilitated the design, with published studies stating mean thresholds for pitch perception from 3 semitones (Kang *et al.*, 2009) to 7.6 semitones (Gfeller *et al.*, 2002), with reported ranges from < 1 semitone to > 25 semitones (Gfeller *et al.*, 2002), although more recently reported upper ranges have been lower, e.g. 6, 8 or 12 semitones (Drennan *et al.* 2010; Kang *et al.* 2009; Jung *et al.* 2010, respectively). Many existing tests were limited by not testing intervals smaller than 1 semitone, only investigating one timbre and only investigating pitch discrimination or pitch ranking ability. The PCT was therefore designed to cater for a wide range of abilities by testing intervals of 0.5, 1, 3, 5, 7 and 9 semitones, using two frequencies, three timbres and investigating both pitch discrimination and ranking ability simultaneously. This research question was therefore addressed in part by the design of the PCT, but was answered by the results with CI users. Whilst this research question really only applies to CI users, the testing of the microtonal version of the PCT with NHL provides a useful additional tool to assess the properties of the PCT. If difficulty levels are similar for the NHL group, then the PCTm when used with NHL may be useful as a model of how CI users interact with the original PCT.

The PCT was shown to provide a suitable, although not perfect, level of difficulty for CI users. In terms of the numbers of individuals affected by ceiling effects, the PCT was comparable to the UW CAMP and the MCI. The PCT discrimination test had 5/22 CI users showing ceiling effects, and the ranking test had 3/22 CI users showing ceiling effects (>80% for all intervals), compared to 3, 5 and 2/22 CI users showing ceiling effects (with scores of 1 semitone) for the UW CAMP 262 Hz, 330 Hz, 392 Hz and 7, 6, 6, 3 and 1/22 CI users showing ceiling effects (with scores of 100%) for the MCI 5, 4, 3, 2, 1 semitones. The presence of ceiling effects in all three tests indicated that they were to some extent too easy for this group of CI users, and a test which shows ceiling effects cannot define ability range accurately within a population, as well as not being able to determine any improvements in ability. The smallest intervals tested by each test were: PCT 0.5 semitones, UW CAMP 1 semitone, MCI 2 semitones (e.g. interval of 1 semitone made up of 2 notes up and 2 notes down, '/\'). Reducing these smallest intervals may increase the difficulty of each test, however this is based on the assumption that CI users' pitch perception ability follows a monotonic psychometric function e.g. that larger intervals are easier and smaller intervals are harder to pitch discriminate or rank, which may not be the case (Levitt, 1971; Looi *et al.*, 2008; Swanson, 2008; Maarefvand, Marozeau and Blamey, 2013). However, for the CI users reported here with ceiling

effects, there was no evidence of non-monotonicity within the intervals 0.5, 1, 3, 5, 7 and 9 semitones, suggesting that for these high performing CI users who were hitting ceiling, reducing the interval in terms of size may remove the problem of ceiling effects.

Numbers of individuals affected by floor effects within the PCT was very low, however this was greater than both the UW CAMP and the MCI. One CI user was affected by floor effects (~50% for all intervals) for the discrimination task and 3/22 CI users for the ranking task (~25% for all intervals), compared to 0, 1, 1/22 CI users who showed floor effects (> 11 semitones) for the UW CAMP (262 Hz, 330 Hz, 392 Hz) and no floor effects (a score of approximately 11%) were seen for any of the MCI intervals. Given the greater difficulty of the ranking task when compared to the discrimination task (Yitao and Li, 2013), the greater number of floor effects for ranking is unsurprising, however the existence of floor effects to a greater extent in the PCT compared to the other tests indicates that it might benefit from additional larger (easier) intervals. It is difficult to say whether it is more important to avoid floor or ceiling effects: ideally, neither would be present, however if it is assumed that pitch perception ability follows the normal distribution, then a test that could encompass any possible ability may have to be extremely varied in terms of its difficulty. Pitch perception ability may not follow the normal distribution due to biological and physiological constraints regarding the smallest perceivable interval (both for NHL and CI users), and so it would be more likely that ceiling effects could be avoided altogether, and some floor effects may need to be accepted as inevitable for some individuals.

The largest interval included in the UW CAMP (12 semitones) was larger than the largest PCT interval (9 semitones), although the UW CAMP still demonstrated some floor effects, suggesting that intervals larger than 12 semitones may be required for some CI users. The UW CAMP tested pitch ranking, and so the requirement to pitch rank 12 semitones may be too hard for some CI users; they may perform better if they only had to discriminate 12 semitones (which is why both tasks were built into the PCT). In addition, the problem of 'octave confusion' (Henry and Meikle, 2000) may play a role in the issues with pitch ranking an interval of 12 semitones, and this was why the interval was avoided in the PCT.

The largest interval that could possibly be utilised as a cue to pitch direction within the MCI was 20 semitones, which was made up of the interval of 5 semitones and the 5 notes of the rising contour ('/'). No floor effects were seen for any interval within the MCI, meaning that no CI user scored as low as 11%, even for the 1 semitone condition (with the rising '/' or falling '\' condition spanning 4 semitones). Each contour contains 5 notes meaning that the 'smallest' intervals are created by 3 notes '/\' and the largest are created by 5 notes'/', '\', much greater than other pitch tests which use only 2 notes to define an interval.

A simpler way to compare the difficulty of these three tests was to look at the overall results in terms of the floor and ceiling effects that were at the boundaries, in either direction, of the tests. As the two parts of the PCT are completed and scored simultaneously, they cannot be performed independently of each other and as such, looking at individual floor and ceiling effects is almost unnecessarily critical: therefore, of interest were any ceiling effects seen at 0.5 semitone for ranking (considered to be the most 'difficult' task), and any floor effects seen at 9 semitones for discrimination (considered to be the 'easiest' task). Similarly, the MCI was never designed to be performed by separate interval. Whilst splitting the intervals up made for ease of programming and presentation to participants, looking at individual floor and ceiling effects can be informative regarding the capabilities of the NHL group, however again this is unnecessarily critical for the MCI. Rather, the focus should be on the floor effects at the largest interval (5 semitones), and the ceiling effects at the smallest (1 semitone). For the UW CAMP, of interest were any individuals showing floor or ceiling effects across every frequency tested. This indicated that the PCT was too easy for 3/22 and too hard for 1/22 CI users. In comparison, the MCI was too easy for 1/22 CI user and the UW CAMP was also too easy for 1/22 CI user, and neither of these tests were considered too hard for any CI users.

In light of what has been discussed earlier regarding the positive skew to an assumed normal distribution, it seems that aiming to design a test that showed no ceiling effects, but some floor effects may be the most desirable goal, as it should be possible to design a test that has intervals small enough to remove all chance of a ceiling effect, however this is not likely to be possible for all floor effects. Therefore, whilst the UW CAMP and MCI appear to have performed slightly better than the PCT, with less ceiling effects and no floor effects, it could be argued that a better result would be no ceiling effects and more floor effects: future proofing the tests for improved ability of CI users and attempting to accurately quantify at least one end of CI users' abilities.

The PCT(m) was also used with NHL, with microtonal intervals. Results indicated that the PCTm provided a suitable level of difficulty for the NHL group, with low levels of floor and ceiling effects for both discrimination and ranking tests. Similar results were seen for the MCIm, indicating that the microtonal intervals chosen for both the PCTm and the MCIm were suitable for this group. Although the approach to measuring pitch perception with these tests was different, the PCTm showed that the range of 5-25 cents was appropriate, and the MCIm, which at its smallest non-flat contour tested either 10 or 20 cents, and at its largest contour tested 20 or 40 cents agreed with these results: this group of NHL pitch discrimination and ranking ability was between 5 and 40 cents. The SOECIC MTB PDT showed no floor or ceiling effects, as its large range encompassed the ability of the NHL.

Both the PCTm and the MCIm can be assessed in terms of ceiling and floor effects that affect the 'easiest' and 'hardest' parts of their tasks. The PCTm's easiest task (discriminating 0.1 semitones) showed floor effects (50% for all intervals) for 2/23 NHL, and its hardest task (ranking 0.05 semitones) showed ceiling effects (≥80% for all intervals) for 2/23 NHL. In comparison, the MCIm's easiest level (0.1 semitones) showed floor effects (11%) for 1/23 NHL and the hardest level (0.05 semitones) did not show any ceiling effects (100%). These results indicate that for NHL, the difficulty levels between the microtonal version of the PCT and the MCI were very similar, with similar results even though the test approaches were so different. A similar comparison was seen in the CI users.

These results suggest that the PCT's difficulty level could be improved to make it even more suitable for testing CI users. Potentially, the upper limit could be improved: an interval larger than 9 semitones may be more inclusive for some CI users that struggle to discriminate between two tones 9 semitones apart. However, as discussed above, there will always be some CI users who will be unable to successfully discriminate between two intervals, regardless of the sizes of intervals chosen. Therefore the acceptance that some floor effects may always be present is important. Rather, the reduction or elimination of ceiling effects may well be possible: there are physical limitations to how small an interval can be that can be successfully discriminated or ranked, and therefore, future designs of this test (and all tests of pitch perception for CI users) should strive towards the removal of ceiling effects. In addition to these points however, it is vitally important that the middle ground between the floor and ceiling effects is not overlooked.

It would be ideal to have a test that stretched to testing intervals as large as 20 semitones, but also tested as small as 0.3 semitones. However, if the interval size of 0.3 semitones was used as a starting point and also as a definition of resolution of the test, there would be 60 musical intervals to test between 0.3 – 20 semitones. This is not feasible for anyone to take part in: using 32 repeats per the 6 intervals took 10 minutes (per condition) for these CI users. Due to the use of the MCS, this would also not be an efficient use of time, and perhaps the original PCT could be used initially, and once an idea of where an individual's area of best performance lies (which may involve more than one interval size), a further, much more detailed version of the PCT could then be used to pinpoint specific ability there.

7.5.3 Does the PCT demonstrate reliability on retest?

In order to compare the reliabilities of the different parts of the PCT, using the ICC, and to then compare these with ICCs from other tests, the PCT's calculated DLs were used. The DL was calculated using a maximum likelihood procedure (described in detail in section 7.3.4), and found

the interval at which the score was equal to 22/32. The calculation of the DL meant that data was lost, because the algorithm sometimes resulted in 'threshold' scores that were outside of a sensible range (e.g. 0-9 semitones), and in addition, at times gave a poor visual fit to the data and was therefore discarded. This resulted in a loss of 19% of the data: the data that was spared therefore did not include extremes in performance, and some non-monotonic psychometric functions were also lost through use of the DL. Any analysis that used the DL data had to take this data loss and therefore bias into account when drawing conclusions (further detailed discussion regarding the use of the DL will be discussed later in section 7.5.9). To try to avoid this problem, an alternative way to compare the differences between T1 and T2 data was used: the average mean differences between T1 and T2 were compared between PCT subtests. This meant that in a very simplistic manner, the smallest differences represented the T1 and T2 that were most similar to one another, and represented the best reliability (details of how this was calculated can be seen in Appendix E). This was only helpful for comparing PCT subtests as these difference scores could not be compared to other test results or the ICC.

As described in Chapter 5, suitable reliability was defined by an ICC of greater than 0.8 ('good to excellent', Pinna *et al.*, 2007), as well as being significantly different from the correlation coefficient expected by chance, using Pearson's critical values of *r* for *n*. For CI users, the PCT met the criteria for 4/6 of the discrimination subtests and 4/6 of the ranking subtests. For NHL, the PCTm met the criteria for 2/6 of its discrimination subtests and 3/6 of the ranking subtests.

Generally, the best reliability from the calculated DL data across both CI users and NHL, was seen for sine tones, with 6/8 of the sine tone subtests (made up of F4 and F5, discrimination and ranking, for NHL and CI users) meeting the criteria, 4/8 of the complex tones meeting the criteria and 2/8 of the piano tones meeting the criteria. It is possible that the sine tone produced the most reliable scores because of the simplistic nature of the tone, and that additional harmonics seen in the complex and piano tones added to the confusion and error with both CI users and NHL. Sine tones did not show the best median results: median scores for the NHL were all very similar, ranging from 0.04 – 0.08 semitones, and median scores for the CI users were much more varied, with sine tones neither showing best nor worst performance for the majority of conditions, except for the ranking score for F5 sine, which had the lowest (best) median score. Another possibility was that because the sine tones suffered data loss due to stimuli issues with CI users, and in addition suffered data loss due to the DL calculation, there were lower numbers of pairs which means that reliability calculations have lower power. Whilst this was the case for the CI users, it wasn't seen for the NHL, and the lowest numbers of pairs did not show a tendency to be the most reliable. If

anything, all the numbers of pairs were on the low side for each ICC, which was likely to result in a low power to detect any significant correlation coefficients.

In comparison, the UW CAMP met the reliability criteria of an ICC of greater than 0.8 and the ICC being significantly different from the critical value of *r* for *n* for one of three of its subtests (392 Hz) and the MCI for all 5 of its subtests, using the same group of CI users. The SOECIC MTB PDT didn't meet the criteria, and the MCIm met the criteria for both 0.1 and 0.05 semitone intervals, using the same group of NHL. This would suggest that the PCT is performing better than both the UW CAMP and the SOECIC MTB PDT in terms of reliability, but that the MCI and MCIm are superior in reliability than the PCT.

Interestingly, the reliability scores for the UW CAMP are very different to results from Experiment 1, where 262 Hz and 330 Hz base notes met this criteria. The only study to report this in the literature states the UW CAMP had an ICC of 0.85, with a sample of 35 CI users (Kang *et al.*, 2009) however it was not specified whether this was for all 3 base notes. Possible reasons for these discrepancies may be related to the small samples tested, in this Experiment 1, 15 CI were tested and in this Experiment 2, 22 CI users were tested, Kang *et al.* tested 35 CI users. It may be that the theoretical problems within the UW CAMP relating to non-monotonic functions in CI users is causing error and therefore discrepancies between T1 and T2. In addition, there does not appear to be a reliable relationship between higher ICCs and certain frequencies across the PCT and the UW CAMP.

The high reliability coefficients of the MCI were also seen in Experiment 1 with CI users, indicating that something about the methodology is obviously very good and extremely reliable with CI users. This may be attributable to the MCS, or it may be due to the redundancy in the stimuli, e.g. each contour uses more than two notes to create intervals and to give cues to the direction of the pitch change. There are no published reports of reliability data for the MCI.

The NHL SOECIC MTB PDT again showed a poor ICC (0.46), similar to the ICC achieved with CI users in Chapter 5 (0.43) but not as poor as the ICC from the first SOECIC MTB PDT reliability study using NHL (of 0.04). The combination of fixed step sizes and mixed intervals (e.g. any given interval may be defined by being above or below the target note) may be the cause of these very poor reliability coefficients.

The ICC comparisons between the PCT using the DL, and the UW CAMP, SOECIC MTB PDT and the MCI for both CI users and NHL have shown comparable performance of the PCT, and potentially superior performance of the PCT over both the UW CAMP and the SOECIC MTB PDT, although not the MCI, in terms of reliability. However, the PCT DL scores were subject to the biased removal of

data due to the issues of calculating the DL, and as such, these conclusions were drawn with caution. To avoid this bias brought about by the calculation of the DL, the average difference between T1 and T2 was calculated, by squaring and square rooting the difference, and then finding the average per interval and averaging these for the overall condition. This provided a dimensionless unit, and so could not be used to compare between the PCT and other tests, however it did allow comparison within the subtests of the PCT (see Appendix E).

The mean differences ranged from 0.77 – 3.99, and so a cut off of 2 was used arbitrarily, in order to draw a line between the better and poorer performing subtests. The smallest differences (< 2) between T1 and T2 were seen for all the F5 subtests, for both discrimination and ranking, and F4 sine discrimination, for CI users. Less of a clear pattern was shown for NHL (F4 piano, F5 sine and complex for discrimination and F5 complex for ranking). It seems clear that the F5 subtests for CI users showed less variation between T1 and T2 – and the median DL scores (although subject to bias) for F5 were better than for F4. This was not seen for NHL, who showed generally similar scores across F4 and F5.

Using the ICC scores to determine reliability, the PCT (using the DL scores) show similar, if not superior reliability when compared to the UW CAMP and the SOECIC MTB PDT, although not as good as the reliability coefficients seen with the MCI, for both CI users and NHL. These scores also indicate that it is the sine tones that seem to be the most reliable: could this be a reflection on the nature of those stimuli (simpler, easier to discriminate and rank?) or is it related to the reduced number of data points within the sine tones due to sine related data loss? In addition, the extremes of performance and some non-monotonic psychometric curves were not included within this analysis, and so the differences between T1 and T2 were also looked at: which showed that for CI users, the F5 differences were much smaller than the F4 differences. In addition, the median DL scores for F5 were much lower (better) than the median DL scores for F4, suggesting that the F5 stimuli were much easier for CI users, and easier tasks lead to less error (from the participant) and more accurate scores.

The poorest performing test in terms of reliability with these two groups of listeners was the SOECIC MTB PDT. The implications of this are that this test, in its current form should not be used, as the measurement error introduced by the tests itself is too great. Mid performing tests such as the UW CAMP with base notes of 262 Hz and 330 Hz should be used with caution, and any conclusions drawn from their use should be done with an awareness of the tests limitations regarding reliability. The use of the UW CAMP is very wide ranging (e.g. Nimmons *et al.*, 2008; Jung *et al.*, 2009; Kang *et al.*, 2009; Won *et al.*, 2010; Jung *et al.* 2012, Maarefvand, Marozeau and

Blamey, 2013, Drennan *et al.* 2015) and so it vitally important to acknowledge these short fallings in it, so that they can be considered when drawing conclusion or making decisions as a result of these papers.

An alternative explanation for the error is that rather than being due to measurement error from the test, it may be error that arose due to the difference in sessions (Hedge, Powell and Sumner, 2017). This seems fairly unlikely in this situation as all second testing sessions were conducted within 3 weeks, with the average being 5.5 days. As all the CI users had at least 11 months experience with the CI, it is very unlikely that they would have experienced any clinical change in this time. The exception to this was CI 10, who only had 3 months experience, although this participant completed T1 and T2 on the same day, thus negating the likelihood of any adjustment occurring between T1 and T2.

The MCI showed superior reliability across every subtest and for both CI and NHL alike. This may have been due to the redundancy of cues within the contours and in addition the use of the MCS.

Implications for the PCT are that some subtests seem to be more reliable than others. The least differences seen between the psychometric functions for CI users were seen for the F5 subtests, and so it seems appropriate that the F5 subtests be kept, and promoted for use with CI users, and the F4 subtests may need to be reconsidered. It is not immediately clear why the F5 subtests performed better than the F4 (although differences in the notes will be discussed in section 7.5.7), and it may be that this area would benefit from further work.

These results indicate that parts of the PCT and the UW CAMP are not reliable enough to be used clinically to evaluate or measure change. In comparison however, the F5 sine and F5 complex conditions of the PCT perform excellently, and are superior to the UW CAMP for reliability.

7.5.4 Does the PCT demonstrate convergent and concurrent validity?

The inclusion of a research question that asked about concurrent and convergent validity was useful in two ways: firstly, it allowed comparison of the PCT's results with the literature regarding known pitch perception ability of both CI users and NHL, in order to provide some validation of the score, and secondly, it allowed comparison with the results from established tests that were conducted at the same time. Validity is often obtained by comparison with the 'gold standard', however, as no gold standard emerged, it was realised that this may not be an appropriate goal. Instead, comparison of PCT results with existing tests served to illustrate the differences and similarities that existed between the PCT and established tests of pitch perception in order to be

informative to researchers and clinicians. This was only possible by using the DL scores from the PCT, and so was subject to the data loss and possible bias described above.

The PCT demonstrated convergent and concurrent validity, for both the CI user and NHL versions. Both the CI user results and the NHL results were similar to previously published data regarding pitch perception ability: the CI data from the PCT showed median discrimination abilities ranging from 0.8 – 3.1 semitones, and the median ranking abilities ranging from 1.5 – 5 semitones. There were wide variations across the subtests, however these were in keeping with previously published work in this area using pitch ranking tests (Gfeller *et al.*, 2002) range: <1->25 semitones; (Nimmons *et al.*, 2008) range: 1-11.5 semitones; (Drennan et al, 2008): 0.6-6 semitones; (Kang *et al.*, 2009) range: 1-8 semitones; (Jung *et al.*, 2009) range: 0.8 – 12 semitones.) These large variations are likely to reflect the huge differences in test methodology, CI user ability and CI technological capability over the years that the studies took place; it is hard to draw meaningful conclusions from these comparisons other than that they are all similarly varied.

The NHL data showed median discrimination abilities ranging from 6-8 cents (0.06-0.08 semitones) and median ranking abilities ranging from 4-7 cents (0.04-0.07 semitones), for frequencies of F4 (349 Hz) and F5 (698 Hz). NHL tend to be similar in their ability to pitch discriminate and pitch rank at frequencies below 4kHz, and Sek and Moore, (1995) demonstrated that at 500 Hz, their 3 listeners were able to pitch discriminate and pitch rank fairly equally, and could hear and rank changes of around 1 Hz, which at 500 Hz equates to a change of 4 cents. Further work showed that not all NHL are able to pitch discriminate and rank equally, Semal and Demany, (2006) found that 3 NHL performed well on both tasks and could discriminate and rank pure tones with DLs of 15 cents (at frequencies ranging between 400-2400 Hz), whereas other NHL required intervals of up to 40 cents to discriminate and up to 317 cents to pitch rank. This led them to postulate that 'frequency shift detectors' are used to pitch rank and either some NHL do not possess them, or that these detectors are not always able to detect such small intervals. Santurette and Dau (2007) showed that 8 NHL had an average just noticeable difference (JND) (for ranking) of 0.6% at 500 Hz, which is 3 Hz, and equates to approximately 11 cents. These studies all support the results of the PCTm, as they indicate JNDs ranging from 4 to 15 cents. The results from the PCTm are therefore in keeping with the results in the literature, which have been obtained with a variety of methods, thus providing evidence of validity and accuracy.

Generally, the PCT results were in agreement with the results from the UW CAMP, SOECIC MTB PDT and the MCI and MCIm. The scores obtained by CI users with the PCT ranged from 0.8 – 5 semitones, with the UW CAMP they ranged from 2 – 5.5 semitones and the scores from the MCI

when using 2, 3, 4 and 5 semitones (with associated usable intervals ranging from 4 – 20 semitones) scored averages of around 75% correct, whilst the MCI 1 (with associated usable interval ranges of 2-4 semitones) scored an average of 55%. This indicated that from a very basic viewpoint, the tests were all producing scores within a very similar range, and in keeping with the previous literature described above.

The PCT ranking scores were also compared statistically with the UW CAMP due to the shared task (pitch ranking) and unit value (semitones). It was found that the median score with base note 392 Hz was poorer than the UW CAMP 262 Hz, F4 complex, F5 complex and F5 piano. This much poorer result for the highest note of the UW CAMP was also previously reported with CI users by Jung *et al.*, (2009), with 12 Korean CI users. They obtained an average of 8.1 semitones (compared to 5.47 here). This pattern was not seen for Nimmons *et al.*, (2008), Kang *et al.*, (2009) or in the current Experiment 1 data. Due to the fact that no pattern was seen with better performance for high or low frequencies, it seems likely that the reason for these differences reflects the diversity of CI users, implants and test methodologies, rather than being caused by the differences in root note.

The scores obtained by NHL with the PCTm ranged from 0.04 – 0.08 semitones, the median score from the SOECIC MTB PDT was 0.16 semitones and the median score from the MCI 0.1 semitone (with a useable interval range of 0.2 – 0.4 semitone) was 44% and from the MCI 0.05 semitone (with a useable interval range of 0.1 – 0.2 semitone) was 25%. Again, from a very basic viewpoint, these tests were all in general agreement regarding the abilities of NHL.

The PCTm discrimination scores were compared statistically with the SOECIC MTB PDT due to the shared task (pitch discrimination) and unit value (semitones). The SOECIC MTB PDT median score was 16 cents (0.16 semitones) and was found to be significantly poorer than the results for all 6 subtests of the PCTm. The literature would indicate that the PCTm is demonstrating greater accuracy here, rather than the SOECIC MTB PDT, and the issues (e.g. fixed step sizes, mixed intervals above and below, use of adaptive procedure, even with NHL) in the SOECIC MTB PDT's approach may be in part responsible for these differences.

The PCT (using DLs) showed results in keeping with the literature for both CI users and NHL, but particularly for NHL, indicating that the PCT methodology and difficulty level were both suitable for assessment of pitch perception in both NHL and CI users, demonstrating convergent validity. In addition, it could be argued that there was some evidence of concurrent validity: the PCT results were shown to be generally similar to results obtained from existing and established tests of pitch perception, when performed on the same population at the same time. Detailed results however showed that there were statistical differences both between the PCTm and the SOECIC MTB PDT,

and between subtests of the PCT and the UW CAMP 392 Hz. This highlights the need for caution when comparing results from two tests claiming to measure the same thing, and emphasizes the importance of transparency in test design and interpretation, in order to avoid inappropriate clinical conclusions and decisions.

7.5.5 Does the PCT demonstrate sensitivity to musicianship?

Known differences within a spectrum can be used to provide validity if they can be demonstrated using a test, e.g. a pitch test should be able to differentiate between those very good at pitch tasks e.g. musicians, and those that are not. All participants were asked whether they considered themselves to be musicians, whether they had any formal music training, and whether they participated in any regular musical activities. Musicianship was defined by answering yes to either of the first two questions, as only including formal qualifications would fail to identify a number of musically able people, and self-report is a common method for this assessment (Law and Zentner, 2012). It was hypothesized that musicians would perform significantly better than non-musicians with the PCT, providing some evidence that the PCT was demonstrating criterion validity.

The PCT did show evidence of sensitivity to musicianship within the NHL group and some indications that it would be sensitive to musicians within the CI group, although with only 3 musicians within this group, statistical analysis was not performed. The only way to compare the PCT scores for musicians and non-musicians was by using the DL, and so the data loss and subsequent bias also affected these results.

Data loss occurred as a result of the PCT DL calculation, and this affected some of the 3 musicians' scores: in the discrimination results (Figure 7.25) only one data point was lost: CI 13 scored so well on F5 piano that the DL calculation attributed a negative score and so this data point was discarded. In the ranking results (Figure 7.26), 4 data points (out of 12) were lost, two were due to high performance leading to a negative score, one was due to non-monotonic/poor performance leading to a score above 9 semitones and one was due to non-monotonic performance leading to a result which was not in keeping with the spread of the data, and so was also discarded (examples are given in Figure 7.24).

The PCT discrimination data showed that musicians were not amongst the poorest performers, however they were not necessarily performing at ceiling. Again, had more musicians been included within the group, there may have been no statistical differences between the groups, as non-musicians performed as well, and in some cases, better than the musicians. The discrimination

data showed very similar results: musicians were not amongst the poorest performers, but still, non-musicians continued to perform as well and better than some musicians.

For both the discrimination and ranking data, there are data points set apart from the other 2 musicians: these were all from CI 15. Whilst it appears that CI 15 was performing the poorest, there were also two data points missing due to poor and non-monotonic data, which were due to CI 13 and CI 17. Also, the performance of CI 15 was around 3 semitones across both discrimination and ranking, which is not a poor performance, and is in keeping with many reported average scores for CI users. On the three occasions across both discrimination and ranking, data loss due to negative data being discarded (e.g. very high performance) was due to CI 13.

As in the PCT, the musician data points within the UW CAMP were spread across the data, and non-musicians performed as well and better than them. For 262 Hz and 392 Hz, CI 15 achieved thresholds of around 4 and 6 semitones, however the poorest score for any of the participants for 330 Hz was achieved by CI 13. The three CI musicians showed excellent performance on the MCI, scoring above 85% for 2, 3, 4 and 5 semitone intervals, and scoring above 75% for 1 semitone. However, many non-musicians also scored at 90% or more, and so even with more musicians, a statistical difference between the groups may not have emerged: this may be due to ceiling effects across both groups within the MCI. The demographic details between the 3 musicians were not noticeably different, aged from 47 – 66 years, and implanted for 18-108 months, with CI 15 (who achieved the 'poorest' musician performance within the PCT data) having been implanted for the longest length of time.

The NHL group had 14 musicians and 9 non-musicians. The PCT discrimination test showed significant differences between these groups for 4/6 subtests: for the sine and complex tones for both F4 and F5. Musicians performed just as well when discriminating the piano tones, however, the non-musicians performed better for piano than for sine and complex, and hence there was no significant difference between musicians and non-musicians. The extra harmonics of the piano tone may have enhanced the non-musicians' performance. Significant differences between the groups were not as frequent in the ranking data: only the complex tones gave musicians the advantage. There were a lot more data points missing for the non-musician data and this may account in part for the lack of significance, and generally, there is a lot more data overlap between the groups for the ranking task. In comparison, both the MCIm and the SOECIC MTB PDT showed statistically significant differences between the groups, and appear to have a much clearer definition between the performances of the musicians vs the non-musicians.

With NHL, the PCT demonstrated evidence of criterion validity, however the lack of (potential) statistical difference between the musicians and non-musicians within the CI group does not

necessarily mean that all 3 tests should be discarded on the grounds of no criterion validity for CI users. These results indicate that other factors are likely to be responsibly and that musicianship alone isn't a strong enough predictor. Does it reflect poor test design in terms of the PCT and UW CAMP when used with CI users? As the PCT shows sensitivity to musicianship with NHL, it seems more likely that test design isn't to blame here, and rather the multitude of factors that affect the success of pitch perception with a CI are likely to be responsible. Unlike the NHL group, CI users are influenced by many more complex factors than the definition of musician, and it may be that experience of being a musician, and music lessons at school play much less of a role than the survival of the spiral ganglion; the success at insertion, and the many other factors that will influence pitch perception success with a CI. For CI users, there is evidence that being a musician isn't enough, instead practicing music prior to and since implantation leads to the best musical results (Maarefvand, Marozeau and Blamey, 2013).

How does this influence use of the PCT? Given the most likely reasons for the lack of significant differences seen in the NHL and CI groups, which are not due to poor test design, all subtests of the PCT should continue to be included on a criterion validity basis.

7.5.6 Are CI users' psychometric functions monotonic?

Whether psychometric functions are monotonic or non-monotonic is of fundamental importance for the methods used in the testing of pitch perception in CI users, as one of the assumptions of the transformed up down method is that the function be monotonic (Levitt, 1971). Adaptive methods for testing pitch perception are therefore flawed (at least for some CI users) and the use of the MCS is argued to be preferable (Swanson, 2008; Cosentino *et al.*, 2016). However, well used and well cited tests still use adaptive methods to assess pitch perception in CI users e.g. the UW CAMP (Nimmons *et al.*, 2008). One of the benefits of the PCT was that the shape of psychometric functions could be estimated because of the use of MCS, and this enabled any non-monotonic functions to be identified. When grading the psychometric functions from the PCT, non-monotonic functions were defined as scores that did not follow either a positive (e.g. best performance when the intervals were larger) or a negative (e.g. best performance when the intervals were smaller) trend across the intervals and were not flat. Pitch reversals were included within the definition of non-monotonicity. A psychometric function was determined as being non-monotonic regardless of whether T1 and T2 both displayed non-monotonic features.

Non-monotonic functions were seen in the CI group in Experiment 2: 55% of CI users had a non-monotonic psychometric function for at least one condition. Of the 428 functions generated, 8%

were non-monotonic. The majority of these were seen for the ranking task, and the majority of these were seen with the piano stimuli.

Non-monotonic psychometric functions have been reported in the literature using many different methods: the adjustment of current to create 'phantom' electrodes has led to non-monotonic functions in pitch ranking ability as the stimulation changes from monopolar to bipolar (Macherey and Carlyon, 2012); electrode discrimination has shown pitch reversals (Kenway *et al.*, 2015); and pitch reversals have been seen in rate pitch studies (Kong and Carlyon, 2010; Cosentino *et al.*, 2016). These studies relate to pitch ranking tasks rather than pitch discrimination, which may just be because ranking tasks are more common and seen as more important than discrimination, and there should be no reason as to why pitch discrimination tasks don't also show non-monotonic psychometric functions in some CI users.

The non-monotonic functions are explained by the many issues with the CI and its interface with the hearing damaged cochlea, specifically the broad current spread, and mismatched frequency placement, as well as issues with cochlea dead regions (Macherey and Carlyon, 2014; Zeng, Tang and Lu, 2014). Many CI users presented a psychometric function with mid performance being the best, and poorer results at the largest and smallest intervals.

The majority of all the non-monotonic functions were seen for the piano stimuli. Previous work using the MCI (Galvin, Fu and Oba, 2008) showed that performance was significantly better with organ stimuli when compared to piano stimuli, due to the spacing of the harmonics. Electrodograms from that study showed that both organ and piano typically stimulated 8-9 electrodes for the note A 440, however for the organ these electrodes were typically adjacent pairs but these pairs were spaced much further away from other pairs when compared with piano. Electrodograms of the MCI at 5 semitones (the largest MCI interval) using piano stimuli showed that electrode 12 (E12) is stimulated for 4 of the 5 notes in the ascending contour: if the listener strongly attended to E12, or E12 had a better interface between electrode and tissue, then this contour may sound flat. Conversely, E12 is only stimulated for 3 of the 5 notes for the MCI piano at 1 semitone; which means that if E12 was used as a strong cue to determine pitch contour, it is likely that a non-monotonic psychometric function would occur.

Non-monotonic psychometric functions that had one or more intervals that scored below chance were classed as containing 'reversals'. This was because scores below chance indicate that the listener is deliberately choosing the opposite of the correct answer, rather than scoring at chance. For pitch discrimination, which has chance levels of 50% and means that 2 of the 4 AFC options are classed as 'correct', reversals meant that the listener repeatedly chose the incorrect pair, at a much more frequent level than if they were performing at chance and were not sure, e.g. instead of

choosing X Y Y, the listener persistently chose Y Y X. There was only one reversal seen for the discrimination task: CI 12, who achieved scores for most intervals ranging from 55-100%, scored around 20% for the 1 semitone interval, at T2 only. Reversal of scores in this way may be caused by a temporal aspect confusion, that the two 'same' notes may have become a single sound, making X Y Y become X Y, thus leading to confusion regarding which sound (e.g. the first or last) was different. This seems unlikely due to the length of the notes and the length of the gap between the notes, and the gap in particular was longer than the gap in the MCI test, as this was felt to be too fast. Rather than a temporal misinterpretation, the task may have been confusing e.g. the listener was unsure how to respond correctly: again this seems unlikely as CI 12 performed well at other intervals and for other stimuli. As this one example of a discrimination reversal was only seen at T2 and not at T1, it suggests that this is not a continuing issue for that frequency and interval combination for CI 12, although the idea that this is due to random noise doesn't seem to make sense as this might suggest that the score would be around chance (50%), rather than being at 24%, (the likelihood of scoring exactly 7/32 by chance alone is 0.078%). If it was due to a lapse in concentration, again it was specific to 1 semitone and not the other intervals, and as each interval was presented at random, it must have been an issue particularly with 1 semitone interval alone.

For ranking, chance level was at 25%, and so scores substantially below this level (e.g. this was set as ~10% or lower) were considered to be a 'reversal', and represented a pitch reversal. There were 12 pitch reversals within the ranking task. For 10 of these reversals, the stimulus was piano, suggesting that when compared to a sine tone and a simple 3 harmonic complex, the additional harmonics within a piano tone may be causing a pitch reversal to be much more likely, as discussed above in relation to the electrodograms from Galvin, Fu and Oba (2008).

Non-monotonic functions were also seen in the NHL group in Experiment 2: 17% of NHL had a non-monotonic psychometric function for at least one condition. Of the 512 functions generated, 1.4% were non-monotonic: 5 from the discrimination task and 2 from the ranking task. There were no pitch reversals within this group. The 4 NHL who showed these non-monotonic functions were non-musicians. In this group, the non-monotonic functions were only seen in the T1 or the T2 version, not both. This indicates that they may be due to individual perception at that time, or lapses in concentration rather than true representations of how the individual NHL perceived that particular interval stimulus combination.

Three of the non-monotonic NHL examples showed a better score for the interval 10 cents, when compared to the other intervals, but only for T1. Perhaps they had higher motivation at T1 and

additionally 10 cents was a pleasing and/or easy interval to discriminate. The other examples included best (and equal) performance of ~75% at 5 and 25 cents at T1 (this good score was similar at T2 for 25 cents only), and for NHL 15 the T1 score for 20 cents was surprisingly poor (~50%) and so perhaps this was a lapse in concentration or lack of motivation, although the fact that it only affected one interval is of interest, because the different intervals were interleaved randomly. The two non-monotonic ranking examples showed generally lower scores than the discrimination task, but best performance at 15 and 20 cents, and then poorer scores as the interval got larger.

Whilst there is no clear reasoning as to why these isolated examples of non-monotonicity occurred for NHL, there are implications that even for some NHL (particularly non-musicians, potentially), adaptive pitch tests may not be appropriate and may converge to some erroneous value, and the danger here is that there is nothing to state that the convergence might be wrong, unlike the methods used in Cosentino *et al.*, (2016). As far as the author is aware, there are no studies looking at non-monotonic pitch perception functions in NHL, and this might be an area for future research, particularly in non-musicians.

The information provided by the PCT in terms of non-monotonic psychometric functions would also be helpful for tailoring music listening advice or practice, both in terms of avoiding, or deliberately targeting areas of pitch reversals or poor performance in order to reach optimal ability (Maarefvand, Marozeau and Blamey, 2013). Since the creation of the PCT by this author in 2011, it has been used in a previous clinical research project to focus on establishing better and worse performing electrodes in order to adjust map parameters and to optimise patient performance and experience, both for musical and general everyday benefit (Grasmeder, 2016, Grasmeder *et al*., 2018).

The results of Experiment 2 using the PCT therefore provide further evidence that CI users and even NHL can show non-monotonic psychometric functions, and although the proportions of CI users (and NHL) who demonstrate these functions are very small, the fact that they exist casts doubt upon the results of adaptive procedures used to test pitch perception in these populations (e.g. Gfeller *et al.*, 2002; Nimmons *et al.*, 2008; van Besouw, 2010; Brockmeier *et al.*, 2011). This evidence does also cast a shadow of doubt over the PCT's use of the DL to calculate a 'threshold', which was used for the statistical analysis throughout this chapter. Whilst this issue has meant that some data has been lost, and subsequent analysis therefore was subject to bias, the psychometric functions that didn't fit the DL calculations were identified and rejected, which is unlike what happens in adaptive procedures such as that used in the UW CAMP, which is an issue that has been raised by Cosentino *et al.*, (2016). Thus, whilst the use of the DL to calculate thresholds is not ideal,

it does appear to have an advantage over the adaptive procedures used by many existing tests of pitch perception for CI users.

7.5.7 Are PCT results affected by stimulus type?

The effects of frequency and timbre were investigated using the PCT. In order to compare average scores from the notes F4 and F5, and from the different timbres: sine, complex and piano, the calculated DL scores from the PCT were used. These were subject to data loss and so the effects of stimulus type are also subject to this bias, as some non-monotonic psychometric functions and calculated DL results that were outside of the limits of the test were rejected. This means that the effects of timbre and frequency on individuals with a non-monotonic function are likely to be lost in this analysis.

Frequency affected the results of the PCT with CI users. The DL scores for the note F5 were significantly better than for F4, for both the discrimination and ranking scores. Previous work using pitch perception tests has not provided overwhelming evidence regarding the effect of frequency in CI users, and the use of the UW CAMP has shown mixed results regarding a most successful base note, with Nimmons *et al.*, (2008) and Kang *et al.*, (2009) showing best average scores with G4 (392 Hz) and Jung *et al.*, (2009) showing best results with C4 (262 Hz). In terms of base notes, the UW CAMP only tested as high as 392 Hz (+ 12 semitones: 784 Hz, G5), making it hard to compare to the highest base note for the PCT, of 698 Hz (+ 9 semitones: 1175 Hz, D6), because the tests only overlapped at the largest intervals of the UW CAMP. Galvin, Fu and Nogaki, (2007) demonstrated much poorer results with the note A3 (220 Hz) compared to A4 (440 Hz) and A5 (880 Hz). Better performance at F5 when compared with F4 was also seen for NHL, for the PCT discrimination scores.

It could be that best performance in any given CI sample might be due to the proximity to best performing electrodes. Without CT scans of the cochlea and details regarding successfully mapped electrodes, this comparison would be hard to do with any level of accuracy, however electrode centre frequencies may provide a starting point regarding why some frequency areas have more success than others. Looking at the cut-off frequencies of the Cochlear Ltd electrodes, the boundary between E21 and E22 is 313 Hz, whereas the boundary between E18 – E19 is 688 Hz, and this boundary is nearer to the note F5 (698 Hz) than the former is to the note F4 (349 Hz). These are the boundary frequencies between electrodes, rather than CFs, and so this information may be unrelated. Current spread and the success of the electrode's interface with the SG cells in the cochlea will also play a huge role and in addition, not all participants were using Cochlear Ltd

devices. In comparison, the data from the UW CAMP obtained with the same group of CI users showed that the best scores were obtained with the base note C4 (262 Hz) and this was significantly better than the note G4 (392 Hz). Performance was better with the lowest of the UW CAMP's notes, and suggests that it isn't just that higher notes result in better performance in this group of CI users, however the PCT and the UW CAMP are very different in their approach to testing pitch perception. The UW CAMP provided a score for every CI user; there was no data loss, however an issue with that was that any data that was not suitable for an adaptive test could not be identified as the test converges and provides a final answer regardless of any potential non-monotonic psychometric function issues that may lead to erroneous convergence (Cosentino, Carlyon, Deeks, Parkinson, & Bierer, 2016; Swanson, 2008).

The PCT data was subject to the loss associated with stimuli issues with the sine tones, plus the use of the DL, and as such, it was wondered whether this data loss could be responsible for creating a significant difference between the notes. For the discrimination data, there was slightly more data loss for F5 than for F4 (F4 had a 29% loss and F5 had a 39% loss). Reasons for the data loss were mixed: there were more data points missing due to DL issues (e.g. outside the test's range) for F4, and there were more data points missing for the F5 data due to participant's only attending on one day (and hence were only taking part in the F4 data). If there had been less data lost, and a larger number of poorer responses for F5, this difference may not have existed. Significantly better scores were also seen for the PCT ranking data, and there was a similar level of data loss between F4 and F5, suggesting that the data loss may not be responsible for this effect. In addition, significantly better scores at F5 were also seen in the NHL group, where there was almost no data loss.

An effect of timbre on PCT scores for CI users was also seen: piano tone scores were significantly better than sine tone scores, in the PCT discrimination scores only. For the easier task of discrimination, the extra harmonics may be providing extra pitch cues, or at least extra differentiation cues (as this effect was for the discrimination task only), when compared to sine tones. No differences in timbre were seen for the ranking task. Whilst there are many studies looking at instrument recognition in CI users, there are very few studies looking at the effect of timbre on pitch. Galvin, Fu and Oba (2008) used different instruments as the MCI stimulus and found significantly better scores with the organ, when compared to the piano. They only tested real instruments, and so there are no sine tones or simple complex tones with which to compare. Electrodograms included within that paper demonstrate how the electrodes of a typical 22 electrode array implant are stimulated for the MCI rising contour at 5 and 1 semitone intervals (Galvin, Fu and Oba, 2008, Figure 1, p191). This demonstrated that for the piano tone many electrodes were stimulated simultaneously, and potentially one of the reasons that the organ

resulted in improved results was that the organ stimulus stimulated electrodes that were more spread out over the array when compared with the piano.

Taken at face value, these results would suggest that complex real-world sounds are more helpful when discriminating pitch, for CI users. These extra cues in the form of harmonics may not be helpful for the more difficult task of pitch ranking, this may be because the complexity becomes more confusing, or it may be due to pitch reversals caused by harmonics. One of the problems with interpreting the data from the PCT is the loss due to the DL calculation, which resulted in the removal of a number of non-monotonic psychometric functions including pitch reversals. Interestingly, of the 9 non-monotonic functions seen for discrimination scores, only 5 were removed by the DL process, and so 4 non-monotonic scores were included in the DL analysis. In contrast, 25/26 of the non-monotonic functions seen for the ranking task were removed by the DL process, and so any analysis using the DL calculation for ranking scores does not take into account the non-monotonic data. This suggests that pitch reversals cannot be responsible for the lack of timbre effect for ranking data; in fact the data included in this analysis is likely to be that of the better performers; so maybe when poorer performers are included, timbre may have more of an effect. This is interesting because a study has shown that musicians are more able to ignore the confusion and illusory effect of timbre in a pitch interval discrimination task when different timbres are used (Zarate, Ritson and Poeppel, 2013). In addition, another study showed that musicians showed a greater advantage than non-musicians when using complex tones, rather than pure tones in an adaptive pitch ranking task (Micheyl *et al.*, 2006). This study also showed that performance was better with complex tones, which was in contrast to Zarate, Ritson and Poeppel, (2013) who found better results with pure tones.

The data loss associated with the discrimination DL (in CI users) may play a role too: data loss was greatest for the sine tones (with a loss of 44%), and the loss for the piano tones was 33%. However, as above, reason for data loss differed across the timbres: the sine tones were mostly lost due to stimulus loudness issues (10 time 1 data points had to be discarded due to unwanted loudness cues, 5 lost due to DL issues and 5 lost due to planned non-attendance on day 2), whereas the piano tones, particularly for F4, were subject to the most data loss due to the DL (10 data points lost due to the DL and 5 data points lost due to planned non-attendance on day 2).

NHL effects of timbre were also seen: the complex tone performed better than the sine and piano for the ranking task only, which is generally in keeping with Micheyl *et al.*, (2006).

The implications that both frequency and timbre have an effect on pitch discrimination and ranking highlights the importance of a test that can measure a range of these features. It also highlights the

need for tests of any kind to be transparent about what stimuli they used as generalisations to other stimuli are not necessarily appropriate. Recognising that frequency and timbre affect pitch perception in CI users (and in NHL) is important both for choice of test (or subtest) but also clinically when interpreting results.

7.5.8 Summary of the PCT

Features of the PCT are summarised in Table 7.11 below. The PCT provided enough repeats and indicated level of success, and also provided a suitable method to test CI users in form of the MCS, by design. A major limitation of the PCT was its smallest interval of 0.5 semitone, indicating that future versions should include a smaller interval, whilst making sure intermediate intervals are not neglected. Floor and ceiling effects, and reliability can be seen for the different subtests, and sensitivity to musicianship (for NHL only) is also displayed.

Table 7.11 Summary of the PCT

PCT subtest	Suitable repeats?	No ceiling effects?		No floor effects?		Small enough interval for CI?	Test retest reliability		Smallest difference (lower half)		Difference between musicians and non-musicians? (NHL only tested)	Suitable (non-adaptive) method for CI?
		NHL	CI	NHL	CI		NHL	CI	NHL	CI		
Discrimination F4 sine	✓	×	×	×	×	×	✓	✓	×	✓	✓	✓
Discrimination F4 complex	✓	×	×	✓	✓	×	×	✓	×	✓	✓	✓
Discrimination F4 piano	✓	✓	×	×	✓	×	✓	✓	✓	×	×	✓
Discrimination F5 sine	✓	×	×	✓	✓	×	×	✓	✓	✓	✓	✓
Discrimination F5 complex	✓	×	×	✓	✓	×	×	✓	✓	✓	✓	✓
Discrimination F5 piano	✓	✓	×	×	✓	×	×	×	×	✓	×	✓
Ranking F4 sine	✓	✓	✓	✓	✓	×	✓	✓	×	×	×	✓
Ranking F4 complex	✓	×	✓	✓	✓	×	✓	✓	×	×	✓	✓
Ranking F4 piano	✓	✓	✓	✓	×	×	×	×	×	×	×	✓
Ranking F5 sine	✓	✓	×	✓	×	×	×	✓	×	✓	×	✓
Ranking F5 complex	✓	×	×	✓	✓	×	✓	✓	✓	✓	✓	✓
Ranking F5 piano	✓	✓	✓	✓	✓	×	×	×	×	✓	×	✓

7.5.9 Limitations to this study

Limitations to this study can be divided into issues with the study design and then issues with design of the PCT. One of the biggest issues within this study was that data was lost due to loudness cues being present within some of the sine tone stimuli. To create the triplets of notes for the PCT, each individual note was generated in Adobe Audition (or Logic Pro, in the case of piano tones), with a mid-level amplitude (to avoid digital peak clipping) that the experimenter predicted would be around 65 dB(A) as presented through the speaker setup in the test room. A Bruel and Kjaer Type 2235 Precision SLM was mounted on a tripod at the position of the listener's ear, although the listener was not present, in order to check that the stimuli were all around 65 dB(A). Particularly with time varying stimuli, it is not easy, and not possible in some cases to create a note that stays at 65 dB(A) across the course of its ringing out. As such, notes were accepted if there loudness was between 60 and 70 dB(A). At the point at which CI 1 – CI 6 had completed the F4 stimuli, and CI 1 – CI 4 had completed F5 stimuli, the experimenter realised that the incorrect stimuli had been presented. These triplets had been recorded incorrectly, and loudness differences, which were large enough for the experimenter to clearly hear were present, meaning that there would be no way to know whether these 6 CI users were performing due to changes in pitch or changes in loudness, and as such the 1st 6 discrimination scores and the 1st 4 discrimination scores for sine tones were removed from analysis.

Not every volunteer CI user was able to attend for 2 separate days, and in order to maximise data collection and to prioritise test retest reliability data points, the experiment was designed so that all of the F4 data was collected on the 1st day, and for CI users that were able to attend for the 2nd day, all of the F5 data was collected then. This was true for 5 of the 22 CI users, and so inevitably there was less data collected for F5, prior to any further data loss due to stimuli issues or DL calculations.

In addition, the typical issues with testing CI users apply to this study: CI users are a rare population, with a diverse range of parameters by definition, different devices fitted over different years, different amounts of experience with the device, and an infinitely diverse set of parameters in terms of implant human interface. As such, sample sizes of CI users are always smaller than desired. Given the diversity of the population, it has been argued that using the median may not be the most sensible way to assess data of populations such as these. The 22 CI users that took part in Experiment 2 used devices from all the main manufacturers, were male and female, and with a relatively diverse age range. However it is unlikely that this sample was truly representative of the CI population in general: this experiment was much more accessible to individuals with time and

ability to travel and attend for 1 to 2 days, which may cause a bias towards unemployed or retired CI users. In terms of musical motivation, this sample may be more representative of highly motivated individuals in terms of musical ability, those that believe that they can improve. Conversely, CI users who have struggled with music listening for a long time and have not seen any improvement may also have been motivated to attend the study. It may not be possible to know how much these factors have influenced the current results, and although power can often be low with small sample sizes, the experimenter recruited as many CI users as possible within the timeframe of June – September 2013.

Use of the ICC

There were also issues with the use of the ICC (A,1). The ICC assumes normal variances within the range of data being compared, and it is designed for use with normally distributed data. These assumptions were not met for all the data within this thesis. One of the main limitations of the ICC is that there are a number of sampling theories on which the wide variations of the ICC are based, and ten different types of ICC (Koo and Li, 2016): not only can this result in the wrong type of ICC being chosen, but results can differ even when the same underlying theory is used, meaning that interpretation is also confusing (Muller and Buttner, 1994, Zhou *et al*, 2011). Significance testing is only well established for comparisons against a correlation of 0, which is lower than the correlations expected by chance for the comparisons in this thesis. The equal variances assumption is easily violated when the ICC is being used to compare a new measure with an existing one. Finally, as with all correlation coefficients, the ICC is affected by the size of the measuring scale: wider ranges result in ICCs that appear to be 'better' (Muller and Buttner, 1994), and a low ICC may be a result of poor agreement, or it may be due to a lack of variability, a small number of subjects or a small number of 'raters' (Koo and Li, 2016).

When the ICC is used, it is essential to its successful interpretation that the associated confidence intervals are reported alongside (Donner, 1986, Koo and Li, 2016), and it has been recommended that when describing the level of reliability, that the confidence interval ranges are used in this way. Koo and Li (2016) suggest that 'ICC values less than 0.5 are indicative of poor reliability, values between 0.5 and 0.75 indicate moderate reliability, values between 0.75 and 0.9 indicate good reliability, and values greater than 0.90 indicate excellent reliability' (Koo and Li, 2016, p158). Therefore, even if an ICC value 0f 0.7 is obtained, if the confidence intervals range from 0.6 – 0.8, this should be reported as a reliability of moderate to good. Whilst the confidence intervals for each ICC were reported within this thesis, they were not used to evaluate the reliability in this way.

A number of things can directly affect the confidence intervals associated with the ICC. Small sample sizes result in a wider confidence interval, and it was been reported that a minimum of *n* =

30 (pairs) be used when using the ICC for reliability analysis (Donner, 1986, Ukoumunne, 2002, Kong and Li, 2016). Throughout this thesis, the largest number of *n* pairs was 22, and often the reliability analysis was lower than this due to some participants not being able to complete all conditions, or due to data loss. As a result, the reliability analysis within this thesis may be underpowered. Low ICCs (and associated wider confidence intervals) can also be caused by low 'rater' (or test) agreement, the lack of variability in the samples, a small number of participants and a small number of raters (or tests) being used (Koo and Li, 2016).

The data sets within this thesis had a mixture of normal and non-normal distributions. The ICC model used here, the ICC (A,1) should be used with parametric data rather than non-parametric data, so this casts doubt upon the conclusions drawn regarding reliability on retest. In addition, the ICC value was compared with the value that might be expected by chance, using Pearson's critical values, and given the presence of non-parametric data, use of Spearman's critical values may have been a more suitable approach.

Significance tests can be used with the ICC to ensure that the value calculated can be considered different from that expected by chance. A significant F test result indicates that the relationship between the two variables is significantly greater than chance (Donner, 1986). The symbol ρ quantifies the level of similarity between the variables (Ukoumunne, 2002), with $\rho = 1$ meaning a perfect correlation and $\rho = 0$ meaning chance level. Normally distributed data is not required for the estimation of ρ, however normality is necessary for the methods used to draw inferences (Donner, 1986, Ukoumunne, 2002). The use of the parametric ICC (A,1) within this thesis could be considered to be inappropriate due to the mixed normal and non-normal data, especially as assumptions within the ICC calculation are heavily dependent on the normal distribution (Donner, 1986). A more appropriate method would have been to use a non-parametric ICC that ranks the data (e.g. Rothery, 1979), however Donner (1986) also states that if the results of the ICC are primarily being used for descriptive purposes, then normally distributed data is not considered so important.

In summary, the use of the ICC within this thesis is questioned due to the small sample sizes and the mixed normal and non-normally distributed data.

Although designed thoughtfully, the creation of the PCT has led to a few issues. One of the important design features of the PCT was the use of the MCS, in order to avoid the issues of assuming CI users have monotonic psychometric functions when it comes to pitch perception. The use of the MCS means psychometric functions can be estimated and visualised, and as a result it was seen that on several occasions and in a large proportion of CI users, non-monotonic

psychometric functions were present. It is difficult, if impossible, to determine one single threshold for pitch perception success with a non-monotonic psychometric function: there may be 2 or more best performing areas across the intervals tested with the PCT. These results cannot be directly compared with the outputs from other tests that create a threshold, e.g. the MedEl MuSIC Test, the UW CAMP, the SOECIC MTB PDT. It is also not easy to assess reliability on retest, or make comparisons between musicians and non-musicians with data that does not provide a single threshold. As such, the MATLAB Palamedes Toolbox was used to fit a psychometric function to the data in order that a threshold result could be calculated: the PCT DL. Creation of the DL allowed statistical analysis of test-retest reliability using the ICC (A,1), it allowed direct statistical comparison with other tests e.g. the UW CAMP and SOECIC MTB PDT, and it allowed statistical comparison between musicians and non-musicians. However, it should be stated very clearly that the use of a single calculated DL goes against what the PCT was designed for: the ability to assess pitch perception ability in CI users with a wide range of abilities, including those with non-monotonic psychometric functions. The calculated DL was used solely for statistical purposes within this experiment, and is not recommended for use clinically.

The MATLAB Palamedes Toolbox fitted a psychometric function to the data using a maximum likelihood procedure, using a logistic function, whilst allowing a 4% lapse rate. The lower point of the curve was set to chance level: this was 50% for the pitch discrimination data, and 25% for the pitch ranking data. The threshold was set to 22/32, which was chosen because a score of 22 or greater meant that the likelihood of this score being due to chance was less than 5%.

The choice of parameters used for the Palamedes function to fit a psychometric curve to the data was done in haste, and as this was an area the experimenter had not had any experience in previously, decisions were led by the existing MATLAB code (fitapf.m). In hindsight, using the logistic function may not have been as sensible a choice as using a curve that was less likely to predict negative values, e.g. a positively skewed distribution curve. Alternative options within the Palamedes Toolbox included the Gumbel or the Weibull distribution curves, and it may be that use of them, rather than the logistic function may have resulted in less data loss.

Another DL issue related to the chance level of 50%: when designing the PCT, and using the Palamedes Toolbox to calculate the DL the experimenter had believed that only one chance cut off could be used, e.g. 22/32, because the PCT assessed discrimination and ranking scores simultaneously. As such, only one threshold level was used within the Palamedes Toolbox, for the calculation of both discrimination and ranking DLs. The cut off of 22 or greater was based upon chance being 50%, and the larger of the 2 chance levels was chosen in order to take the more conservative route, as using the chance level of 25% for both pitch discrimination and ranking

would lead to error. It wasn't until much later in the write-up, that the experimenter realised that the threshold cut off for DLs for pitch ranking could have, and should have been calculated using a chance level of 25%, which would make the cut off 13/32 rather than 22/32. Essentially, this is an issue of statistical analysis of the PCT ranking data only. It does not affect the clinical utility of the PCT. This issue is unlikely to have affected any test retest reliability analysis, as the actual score is relatively unimportant for this analysis. There may have been a different distribution of data loss as a result of this issue, although it is not known whether the data loss would be greater or less. This issue may have a more pronounced effect on the analysis involving musicians and non-musicians, and of the comparison with other tests. If the PCT ranking data was recalculated using 13/32 as a threshold cut off, it is expected that the DL thresholds would be reduced, thus providing an improved average score. This may result in the PCT ranking scores being better than other tests and leading to more significant differences e.g. between the PCT ranking and the UW CAMP. Significant differences were seen between musicians and non-musicians using PCT data, however these differences may extend to a greater number of subtests of the PCT.

As well as causing issues with non-monotonic psychometric functions, other issues arose as a result of using the DL: the MATLAB function occasionally produced extreme scores: very good performers who scored at or near ceiling for every interval were often predicted a negative value interval size for the point at which it was estimated they would score 22/32. In addition, very poor performers were often predicted very large interval sizes (e.g. 4000 semitones) for the point at which it was estimated they would score 22/32. Neither of these score types were accurate, or indeed possible and they had to be discarded. Any scores that fell on the negative side of 0 were discarded. It was more difficult to decide on an upper bound for PCT cut off scores, however it was decided that 9 semitones should be the upper accepted score, as this was the limit of intervals tested within the PCT. Conversely, the cut off point for the lower limit was not set to 0.5 semitone, as it was felt that this would discard too much useful data, especially in the case where participants were performing so well that the Palamedes estimate was helpful and likely to be predictive, rather than being nonsense. However, every result of less than 0.5 semitone was visually checked to ensure that the interval threshold as calculated by the DL was in keeping with the general levels of success of that participant for each condition. An issue with this is that this is different to the way the experimenter treated the raw data from the UW CAMP, which may have put the PCT at an unfair advantage.

The data loss as a result of discarding negative scores and scores above 9 and below 0.5 semitone if they visually did not make sense, meant that analysis using DLs were subject to bias. Some data from very good and very poor performers was removed, and a large proportion of non-monotonic

psychometric functions, particularly for ranking data, were also discarded. This did not however affect the analysis of whether there were enough repeats, nor the analysis of difficulty level of the PCT. To try and get around this issue with regards to reliability, an alternative measure of reliability was utilised, using the raw data scores rather than the DL.

Further work is necessary to establish the Minimum Clinically Important Difference (MCID) (Vaz *et al.*, 2013) which would enable clinicians to determine whether a change in score should be deemed as clinically significant.

Chapter 8 General Discussion

8.1 Are the existing measures of pitch perception used in this thesis suitable for CI users?

From a perspective of establishing a baseline ability level, and being able to assess the magnitude of any change, none of the existing pitch perception tests presented within Experiment 1 of this thesis, in their current state, are ideally suited for this use with CI users.

There are issues with difficulty level: only the SOECIC MTB PDT and 2 intervals of the MCI showed no ceiling and no floor effects, the remainder of the tests demonstrated issues with at least floor or ceiling effects. There are also issues with test-retest reliability, with only 2/3 of the base notes in the UW CAMP, and (all of) the MCI subtests showing suitably high correlation coefficients.

The use of adaptive procedures is a cause for concern with CI users as it cannot be assumed that their underlying psychometric curve for pitch perception is monotonic, which is a requirement of the adaptive staircase procedure (Levitt, 1971). This means that the approach used by the MedEl MuSIC Test, the UW CAMP and the SOECIC MTB PDT may mean that erroneous results are obtained with CI users that have a non-monotonic psychometric function for pitch perception (Swanson, 2008), and questions their validity when being used with CI users.

There are also issues regarding the clarity of what each test is actually testing e.g. pitch discrimination or pitch ranking, and for CI users that may be able to discriminate pitches but cannot successfully pitch rank, tests of pitch ranking only will not be appropriate for providing suitable feedback about their level of ability.

Of the 7 existing tests of pitch perception for CI users presented in this thesis, none meet all the criteria of desirable characteristics wanted in a modern pitch perception test for CI users:

- a suitable level of difficulty
- a sufficient number of repeats
- reliability on retest
- use of a non-adaptive method
- clear interpretation of results for both clinician and CI user

8.2 Is there a best performing pitch perception test for CI users?

The best performing test appears to be the MCI: it shows minimal floor and ceiling effects when used with CI users, it uses a non-adaptive method to collect data, and it shows high reliability across all 5 intervals. However it is not without its problems: the 9 different contours represent different levels of complexity, with the simplest 'contour' being the flat, and the most complex being rising-falling or falling-rising, because they consist of 2 different directions within one contour. Some CI users may not be able to pitch rank successfully enough to choose the correct contour, however may be confident that they heard the flat 'contour', which adds complexity to calculating the role of chance in this test. Galvin *et al.* (2007) found that the flat contour was responded to most frequently and the falling contour the least frequently, indicating that the rising and falling contours may be different in terms of their difficulty. The test would benefit from a specific number of repeats being required when used clinically in order to give statistical confidence in the final result, and in addition interpreting the results is complicated by the fact that each contour (with the exception of the flat contour) spans at least 2 and at most 4 intervals. This means that there may be redundancy in the pitch perception cues.

8.3 Is the PCT an improvement on existing tests and is it suitable for use with CI users?

The PCT shares a lot of its test design with the MCI, however it is considered to be an improvement on the MCI because the complexity is removed, as the PCT only uses 4 contours which are much more equal than the 9 contours of the MCI. Each of the 4 contours within the PCT are equal in terms of complexity, however if rising and falling are subject to different perceptual biases, then these 4 contours may not be equal. The gaps between notes are larger than in the MCI, with the aim of making it less challenging for CI users. The PCT provides a large number of repeats, and gives guidance regarding how many correct trials are needed in order to have confidence that any given interval was successfully pitch discriminated or pitch ranked. The simpler approach to contours means that 2 tones only define the interval being tested e.g. the 2 semitone condition is testing 2 semitones.

The PCT uses a novel way to test the pitch discrimination and pitch ranking ability simultaneously, giving it an advantage over the existing tests of pitch perception for CI users. It allows estimation of the psychometric curve, again, something the existing tests presented in Experiment 1 cannot do. It provides enough repeats to ensure that the effect of chance is low, once a certain level of

success is achieved: this also allows for lapses in concentration. It shows similar numbers of floor and ceiling effects to the UW CAMP, although has more floor and ceiling effects than the MCI. The PCT shows good performance on test retest reliability: it meets the reliability criteria for more subtests than the UW CAMP and the SOECIC MTB PDT, although the MCI's reliability is superior to the PCT. The results of the PCT are in keeping with published literature for both CI users and NHL, and are generally in keeping with the results obtained from existing tests of pitch perception taken at the same time, subject the wide variation seen in CI users. The PCTm, when used with NHL showed sensitivity to musicianship, more so for the discrimination than for the ranking task.

One of the most important advantages of the PCT is its use of the MCS rather than adaptive measures, allowing psychometric functions of pitch perception ability to be estimated. Tests that use MCS are suitable for CI users with non-monotonic psychometric functions. From a clinical perspective, the PCT provides information about pitch perception ability across the breadth of the intervals tested, rather than focusing on a final 'threshold' score. This means detailed feedback can be provided to the CI user, or more practically, it is possible that alterations can be made to the device. The PCT showed floor and ceiling effects, and so the difficulty level currently limits its utility in establishing a full range of baseline measures in the current group of CI users.

In summary, the PCT is considered to be an improvement on existing tests in the following ways:

1. Ability to test pitch discrimination and pitch ranking simultaneously
2. Use of the non-adaptive MCS allowing CI users with non-monotonic psychometric functions to be tested and allows the estimation of the psychometric curve
3. Provides enough repeats to allow statistical confidence in the results and specifies levels needed for success
4. Provides a level of difficulty similar to that of the UW CAMP (although the PCT is limited by ceiling effects in its current state)
5. Is sensitive to musicianship in NHL
6. Shows similar reliability on retest to the MCI and superior reliability to the UW CAMP

The F5 sine tone and complex tone showed excellent reliability on retest, and showed the lowest average differences between T1 and T2, making them the most reliable subtests of the PCT. In addition, they both showed sensitivity to musicianship in NHL, and F5 sine and F5 complex are

considered the most robust subtests of the PCT, and are recommended for use with the PCT going forwards.

8.4 What are the limitations of the PCT?

The PCT is not without its limitations. Due to the use of the MCS, completing the PCT can be time-consuming.

There are issues surrounding the use of a DL; the PCT was designed to provide information about CI users' abilities across different intervals, and did not want to produce an end result that was a threshold, or a DL. As such, the PCT produces a percentage level of success for each interval it measures, with a suggesting indication that scoring over a certain amount (22/32 for pitch discrimination and 13/32 for pitch ranking) indicates success. In addition, the PCT allows DLs to be calculated (outside of the PCT software) in order to determine a threshold which allows for statistical analysis and comparison with other 'threshold' based tests. However in order to make comparisons with existing tests and published literature, and to compare musicians and non-musicians, and to establish test-retest reliability, a DL was required. The calculation of this DL meant that data was lost: data that fell outside the limits of the test, data that didn't appear to visually agree with the DL results, and some non-monotonic data was also lost. This data loss may have been minimised with an adjusted approach to calculating the DL: rather than fitting a logistic curve, a more positive skewed curve might have resulted in less data loss. Extending the limits of the test may result in less data loss, however this is an area for further work in order to know how far from the original intervals tested results can be extrapolated. Unlike adaptive tests, where any error in asymptote goes unnoticed (Cosentino *et al.*, 2013), the PCT is able to visualise and therefore determine non-monotonic psychometric functions and therefore it has an advantage over adaptive tests.

In its current form, the PCT requires a score of 22 or more out of 32 in order to be confident that that interval was heard correctly. It was discovered that this cut-off point of 22 was suitable for discrimination data only, and that for ranking data this cut-off point should actually have been 13/32, because the reduced level of chance (25%) due to the 4 alternative choices meant that the likelihood of this being due to chance was reduced. As such, the results from the ranking data are likely to be improved, if the data was recalculated.

As in the MCI, the PCT also makes the assumption that rising and falling tones are equal in terms of difficulty and salience, which may not be the case, as indicated by Galvin *et al.*'s (2007) reference to the unequal responses between flat contours and falling contours.

The PCT has shown different levels of reliability between F4 and F5, for both CI users and NHL, and it is unclear as to why, as they both use the same methodologies. As such the tests with the highest reliability (e.g. the F5 subtests) are recommended to be used from this point forward, and further work in this area may explain these differences.

The PCT does not currently provide equally spaced intervals with which to test.

The PCT DL calculation cuts off anything above 9 semitones, but accepts DLs calculated that are below 0.5 semitones subject to visually agreeing with the data, and the validity of these decisions are not yet known.

8.5 Future directions

Based upon the results of experiment 2, it is recommended that the PCT be used with the F5 stimuli only from this point forwards, as it shows greater reliability.

In order to address the issues with difficulty level and the existence of ceiling effects, it is proposed that the next stage of the PCT is to incorporate smaller intervals that are equally spaced across the current breadth. The PCT in its current form could be used with a reduction in the number of trials (and stimuli) in order to get an idea of ability, then a more detailed stage of the PCT would take place. The more detailed stage would have 3 levels, and details can be seen in Appendix F.

Issues with this relate to semitone thirds not being recognised as regular musical intervals. However this approach may provide a much more detailed level of pitch perception ability for both discrimination and ranking for CI users, and may reduce the level of ceiling effects.

The implications of this work relates to providing information about a CI users baseline ability, and this can be achieved with different timbres, which potentially allows for advice regarding instrument choice.

The PCT has also been used by Mary Grasmeder, a clinical scientist in USAIS to determine whether patients would benefit from altering electrode frequency allocations or whether there is any benefit to switching electrodes.

8.6 Conclusions

The main conclusions from this thesis are:

1. Existing tests of pitch perception may not be suitable for accurately assessing the pitch perception of CI users.

2. Pitch perception tests for CI users should include enough repeats to ensure that results cannot be attributed to chance; they should not demonstrate floor or ceiling effects; they should demonstrate high levels of reliability; be sensitive to known differences of pitch perception ability (e.g. comparing musicians and non-musicians) and should use methods that are suitable for use with CI users (e.g. should avoid the use of adaptive procedures).

3. The PCT showed superior performance to existing tests of pitch perception in terms of the number of trials presented; the PCT was superior in terms of assessing both pitch discrimination and pitch ranking within the same test and simultaneously; and the PCT was the only test of pitch perception that allowed the psychometric function to be estimated and therefore was the most suited test to individuals who may demonstrate a non-monotonic psychometric function for pitch perception. The PCT showed similar performance to the UW CAMP in terms of reliability on retest, and similar performance to both the MCIm and the SOECIC MTB PDT with regards to sensitivity to musicianship within the NHL group. The PCT showed both floor and ceiling effects with CI users and so this an area that needs improvement. The subtests F5 sine and F5 complex were the most reliable and sensitive to musicianship in NHL and so these are recommended to be the most informative subtests within the PCT.

4. Non-monotonic psychometric functions were seen in both CI user and NHL populations in this thesis. This has implications for the suitability of the use of adaptive procedures when assessing pitch perception in both CI users and NHL.

5. Future work to improve the accuracy of the PCT should be considered to obtain a greater level of information regarding CI users' abilities at perceiving pitch discrimination and ranking at a greater number of intervals that are smaller than one semitone. This may be achieved by dividing the test into two parts, the first which assesses a wider range of intervals and the second, more detailed test, which uses more trials which are cantered around particular interval ranges of interest.

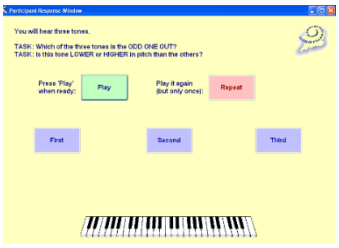
Appendix A

Participant instructions *Test order is randomised*

<table>
<tr>
<td>

SOECIC music test battery

~ 5 minutes

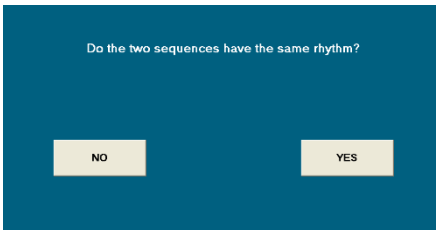


1. Press green play button
2. Choose the odd one out
3. Was the odd one higher or lower? (pitch only)
4. Do not repeat

</td>
<td>

MACarena

~ 4 minutes / 25 trials



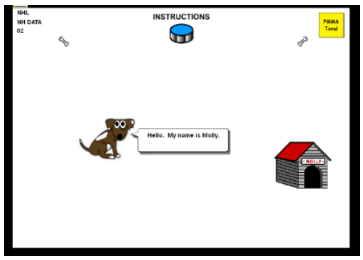
1. Are they the same?

</td>
</tr>
<tr>
<td>

PMMA

~ 12 minutes / 40 trials

designed for children



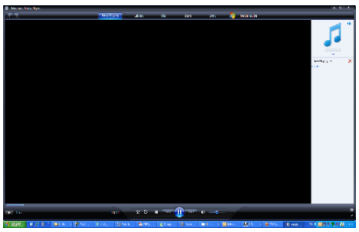
1. Press green play button

Same or different☺☺or☺☹

</td>
<td>

MBEA

~ 10 minutes / 31 trials

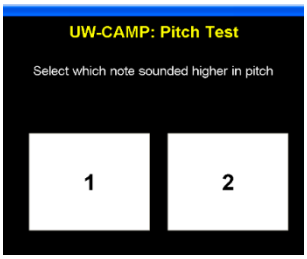


1. Warning beep indicates start
2. Say out loud: same or different

</td>
</tr>
<tr>
<td>

UW CAMP

~ 6 minutes

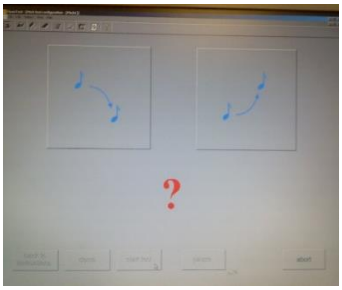


1. Which sounds higher, 1 or 2?

</td>
<td>

MedEl MuSIC Test

~ 6 minutes

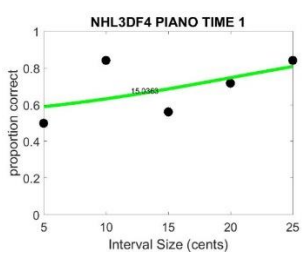


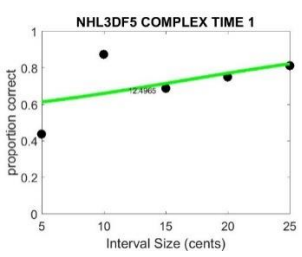
1. Choose the correct box

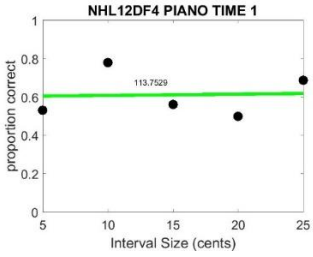
</td>
</tr>
</table>

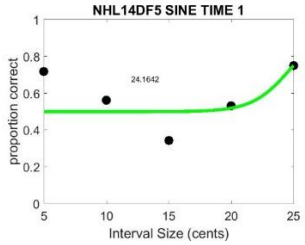
Appendix B

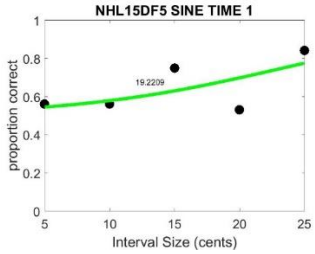
Non-monotonic psychometric functions from NHL

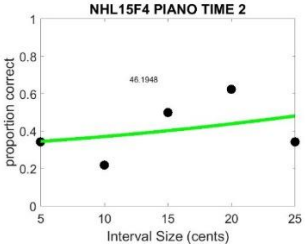


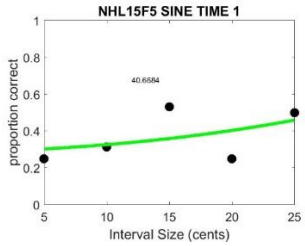






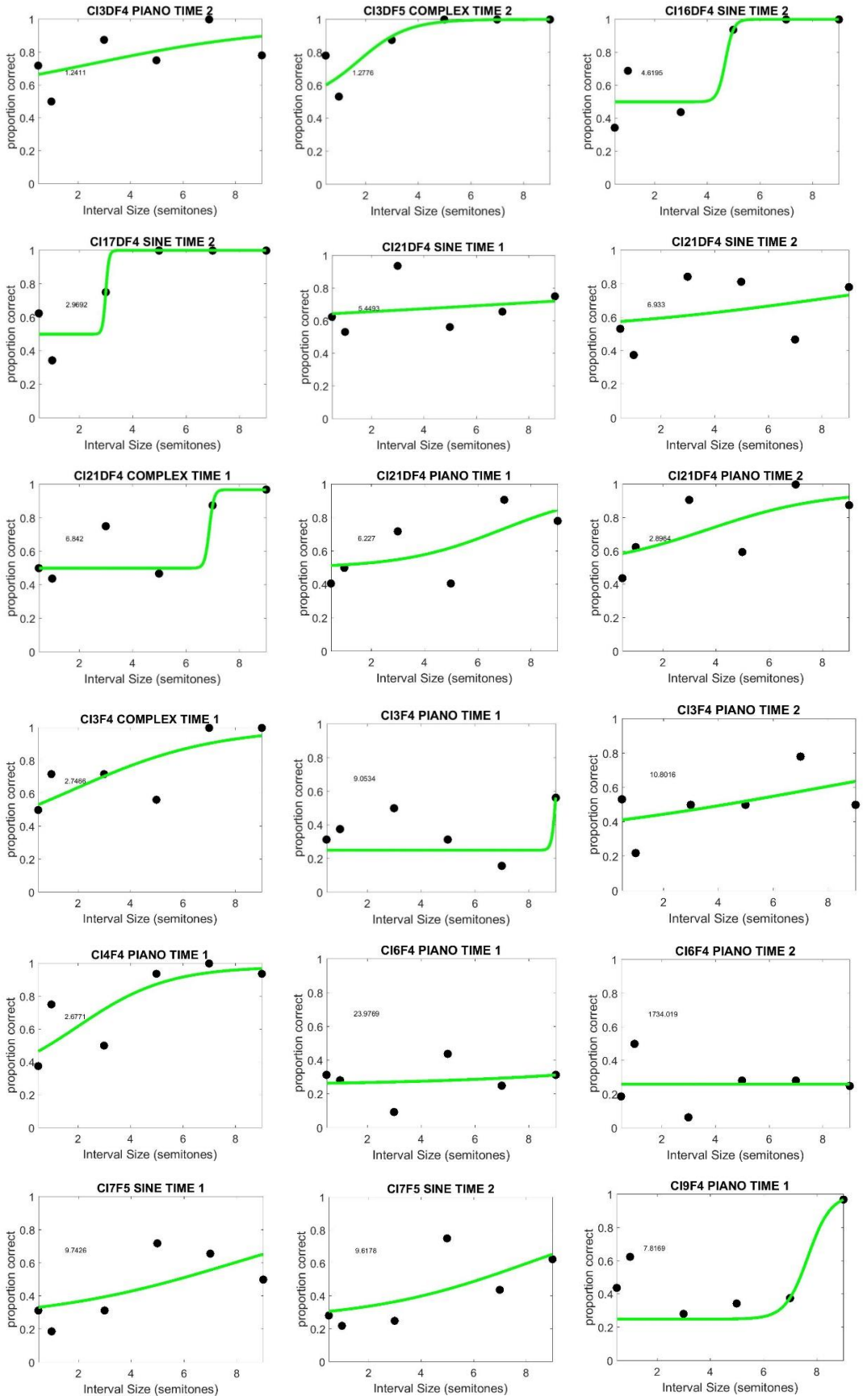


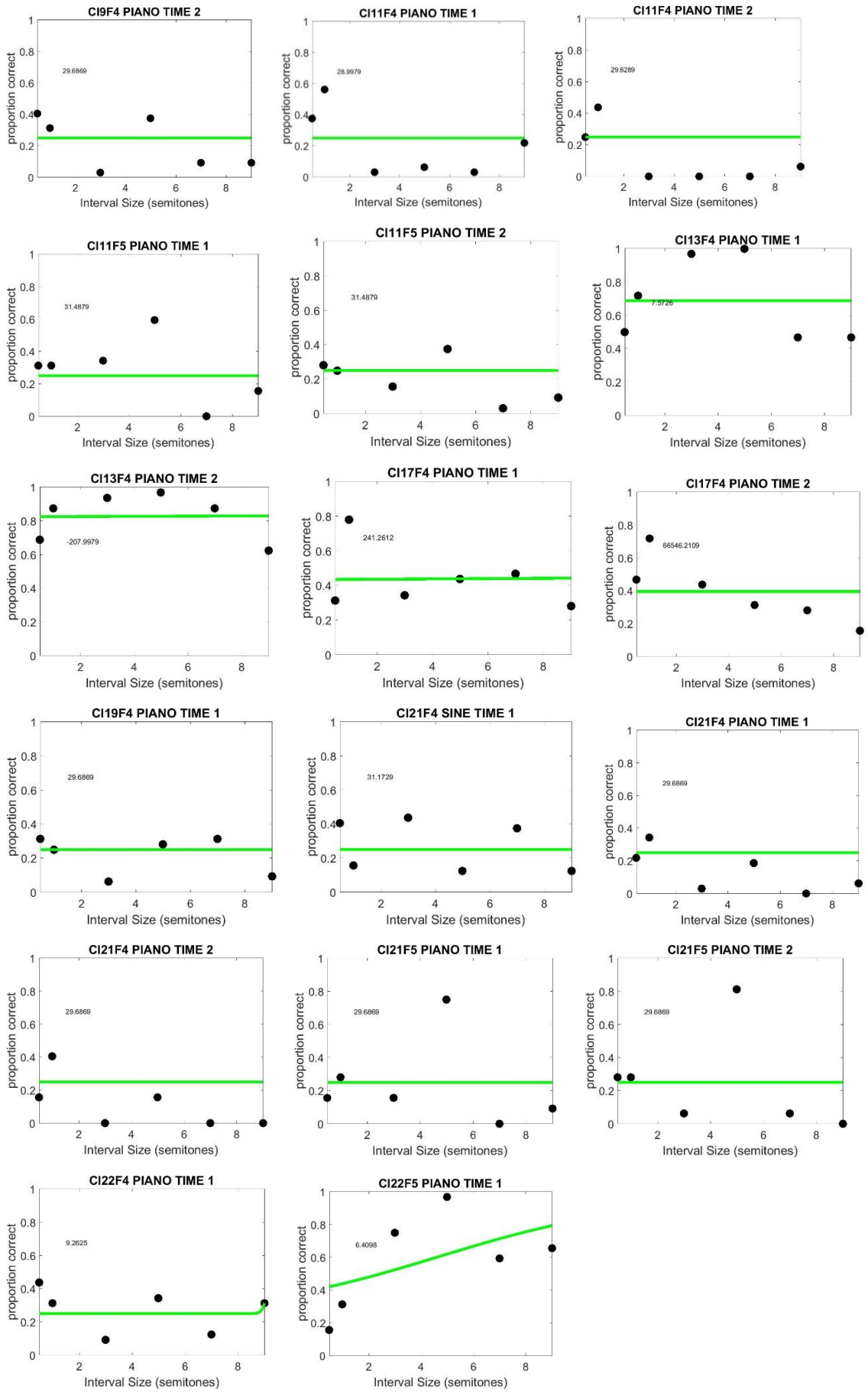




Appendix C

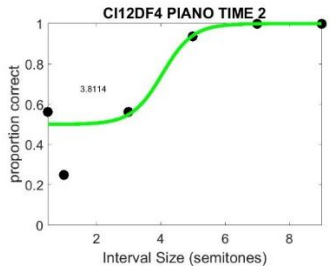
Non-monotonic psychometric functions from CI users



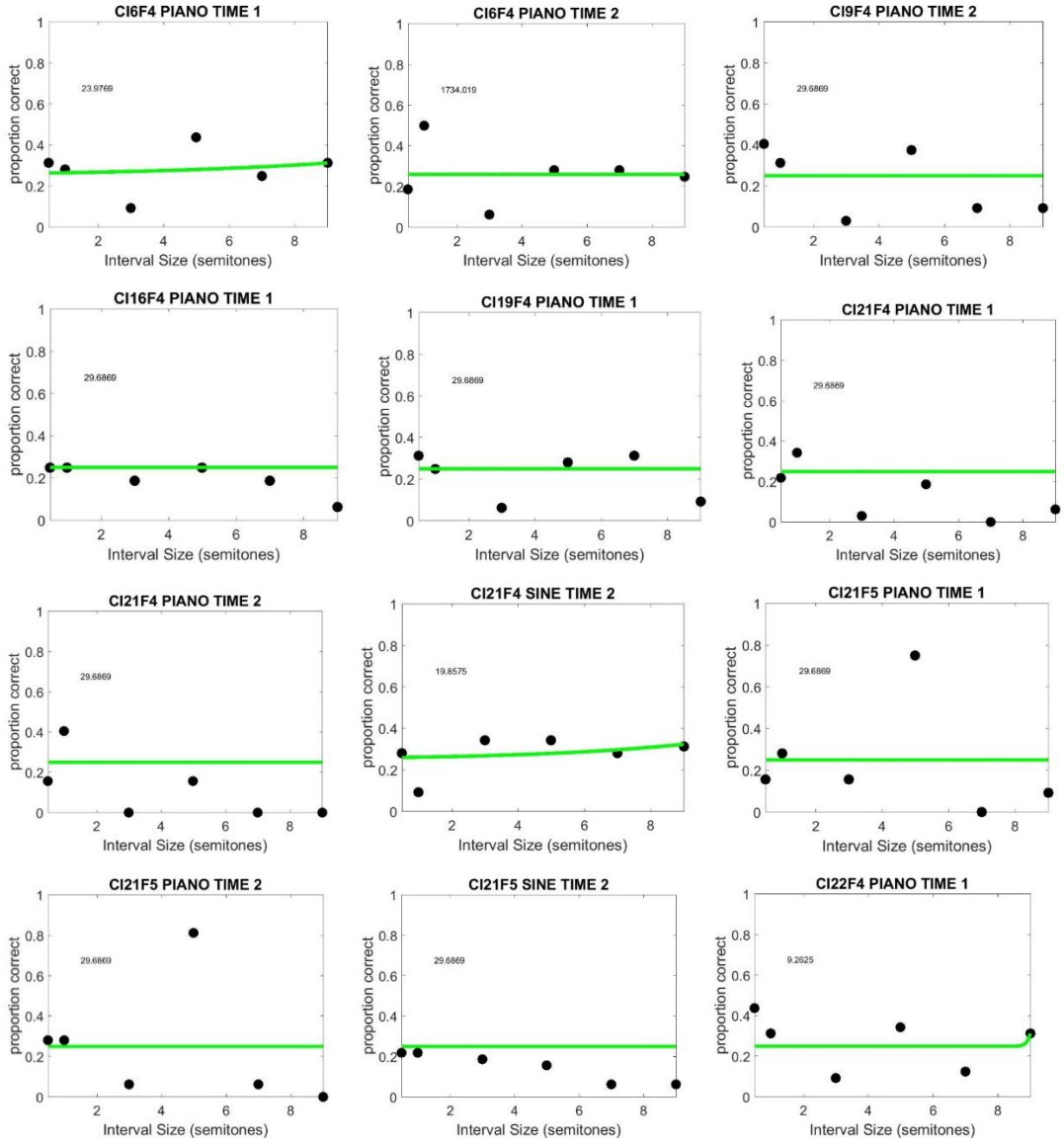

CI9F4 PIANO TIME 2
29.6869
CI11F4 PIANO TIME 1
28.9979
CI11F4 PIANO TIME 2
29.6289
CI11F5 PIANO TIME 1
31.4879
CI11F5 PIANO TIME 2
31.4879
CI13F4 PIANO TIME 1
7.5726
CI13F4 PIANO TIME 2
-207.9979
CI17F4 PIANO TIME 1
241.2612
CI17F4 PIANO TIME 2
66546.2109
CI19F4 PIANO TIME 1
29.6869
CI21F4 SINE TIME 1
31.1729
CI21F4 PIANO TIME 1
29.6869
CI21F4 PIANO TIME 2
29.6869
CI21F5 PIANO TIME 1
29.6869
CI21F5 PIANO TIME 2
29.6869
CI22F4 PIANO TIME 1
9.2625
CI22F5 PIANO TIME 1
6.4098
proportion correct
Interval Size (semitones)

Appendix D

PCT 'reversals': lower than chance score for PCT discrimination



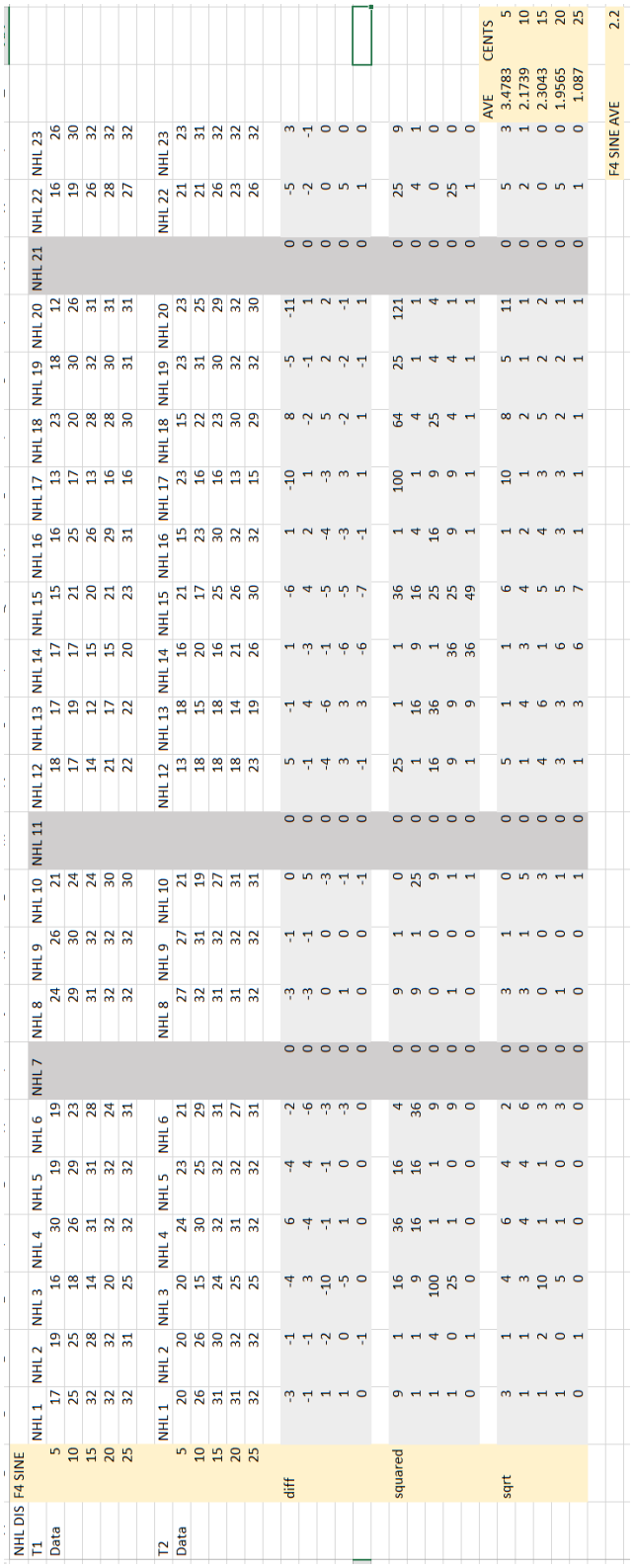
PCT 'reversals': lower than chance score for PCT ranking



Appendix E

The alternative to the ICC used to calculate differences between T1 and T2 data for the PCT.

Example shown is for NHL with F4 sine.



NHL DIS	F4 SINE	NHL 1	NHL 2	NHL 3	NHL 4	NHL 5	NHL 6	NHL 7	NHL 8	NHL 9	NHL 10	NHL 11	NHL 12	NHL 13	NHL 14	NHL 15	NHL 16	NHL 17	NHL 18	NHL 19	NHL 20	NHL 21	NHL 22	NHL 23
T1																								
Data	5	17	19	16	30	19	19		24	26	21		18	17	17	15	16	13	23	18	12		16	26
	10	25	25	18	26	29	23		29	30	24		17	19	17	21	25	17	20	30	26		19	30
	15	32	28	14	31	31	28		31	32	24		14	12	15	20	26	13	28	32	31		26	32
	20	32	32	20	32	32	24		32	32	30		21	17	15	21	29	16	28	30	31		28	32
	25	32	31	25	32	32	31		32	32	30		22	22	20	23	31	16	30	31	31		27	32
T2		NHL 1	NHL 2	NHL 3	NHL 4	NHL 5	NHL 6		NHL 8	NHL 9	NHL 10		NHL 12	NHL 13	NHL 14	NHL 15	NHL 16	NHL 17	NHL 18	NHL 19	NHL 20		NHL 22	NHL 23
Data	5	20	20	20	24	23	21		27	27	21		13	18	16	21	15	23	15	23	23		21	23
	10	26	26	15	30	25	29		32	31	19		18	15	20	17	23	16	22	31	25		21	31
	15	31	30	24	32	32	31		31	32	27		18	18	16	25	30	16	23	30	29		26	32
	20	31	32	25	31	32	27		31	32	31		18	14	21	26	32	13	30	32	32		23	32
	25	32	32	25	32	32	31		32	32	31		23	19	26	30	32	15	29	32	30		26	32
	diff																							
		-3	-1	-4	6	-4	-2	0	-3	-1	0	0	5	-1	1	-6	1	-10	8	-5	3	0	-5	3
		-1	-1	3	-4	4	-6	0	-3	-1	5	0	-1	4	-3	4	2	1	-2	-1	-1	0	-2	-1
		1	-2	-10	-1	-1	-3	0	0	0	-3	0	-4	-6	-1	-5	-4	-3	5	2	2	0	0	0
		1	0	-5	1	0	-3	0	1	0	-1	0	3	3	-6	-5	-3	3	-2	-2	-1	0	5	0
		0	-1	0	0	0	0	0	0	0	-1	0	-1	3	-6	-7	-1	1	1	-1	1	0	1	0
	squared																							
		9	1	16	36	16	4	0	9	1	0	0	25	1	1	36	1	100	64	25	121	0	25	9
		1	1	9	16	16	36	0	9	1	25	0	1	16	9	16	4	1	4	1	1	0	4	1
		1	4	100	1	1	9	0	0	0	9	0	16	36	1	25	16	9	25	4	4	0	0	0
		1	0	25	1	0	9	0	1	0	1	0	9	9	36	25	9	9	4	4	1	0	25	0
		0	1	0	0	0	0	0	0	0	1	0	1	9	36	49	1	1	1	1	1	0	1	0
	sqrt																						AVE	CENTS
		3	1	4	6	4	2	0	3	1	0	0	5	1	1	6	1	10	8	5	11	0	5	3	3.4783	5
		1	1	3	4	4	6	0	3	1	5	0	1	4	3	4	2	1	2	1	1	0	2	1	2.1739	10
		1	2	10	1	1	3	0	0	0	3	0	4	6	1	5	4	3	5	2	2	0	0	0	2.3043	15
		1	0	5	1	0	3	0	1	0	1	0	3	3	6	5	3	3	2	2	1	0	5	0	1.9565	20
		0	1	0	0	0	0	0	0	0	1	0	1	3	6	7	1	1	1	1	1	0	1	0	1.087	25
																							F4 SINE AVE	2.2

Appendix F

Suggested intervals for future design of the PCT

0.33 to 3 semitones

Intervals tested: 0.33, 0.66, 1.0, 1.33, 1.66, 2.0, 2.33, 2.66 and 3.0 semitones.

3.33 to 6 semitones

Intervals tested: 3.33, 3.66, 4.0, 4.33, 4.66, 5.0, 5.33, 5.66, 6.0 semitones

6.33 to 9 semitones

Intervals tested: 6.33, 6.66, 7.0, 7.33, 7.66, 8.0, 8.33, 8.66, 9.0 semitones.

List of References

ANSI (1994) American National Standard Acoustical Terminology. American Standards Institute. New York

Arnoldner, C. *et al.* (2007) 'Speech and music perception with the new fine structure speech coding strategy: preliminary results.', *Acta oto-laryngologica*, 127(12), pp. 1298–1303. doi: 10.1080/00016480701275261.

Arora, K., Dowell, R., Dawson, P. (2012) 'Cochlear Implant Stimulation Rates and Speech Perception' In *Modern Speech Recognition Approaches with Case Studies* Edited by S. Ramakrishnan

Assmann, P. F. and Summerfield, Q. (1990) 'Modeling the perception of concurrent vowels: Vowels with different fundamental frequencies', *Journal of the Acoustical Society of America*, 88(2), pp. 680–697.

Bachem, A. (1950) Tone height and tone chroma as two different pitch qualities, *Acta Psychologica*, 7, pp. 80–88.

Bachem, A., (1937) Various Types of Absolute Pitch. *The Journal of the Acoustical Society of America,* 9, 146

Bartlett, J. W. and Frost, C. (2008) 'Reliability, repeatability and reproducibility: Analysis of measurement errors in continuous variables', *Ultrasound in Obstetrics and Gynecology*, 31(4), pp. 466–475. doi: 10.1002/uog.5256.

Başkent, D. and Shannon, R. V. (2004) 'Frequency-place compression and expansion in cochlear implant listeners', *The Journal of the Acoustical Society of America*, 116(5), p. 3130. doi: 10.1121/1.1804627.

Bernstein, J. G. W. and Oxenham, A. J. (2006) 'The relationship between frequency selectivity and pitch discrimination: Sensorineural hearing loss', *The Journal of the Acoustical Society of America*, 120(6), pp. 392 Hz 9–3945. doi: 10.1121/1.2372452.

Bernstein, J. G., Stakhovskaya, O. A., Jensen, K. K., (2018) Measuring spectral asymmetry for cochlear-implant listeners with single-sided deafness. *The Journal of the Acoustical Society of America* 143, 3

van Besouw, R. M. (2010) 'SOECIC MTB Installation guide'. Available at:

http://www.ncbi.nlm.nih.gov/pubmed/18179057.

van Besouw, R. M. and Grasmeder, M. L. (2011) 'From TEMPO+ to OPUS 2: what can music tests tell us about processor upgrades?', *Cochlear implants international*, 12 Suppl 2, pp. S40-3. doi: 10.1179/146701011X13074645127513.

Bierer, J. A. (2007) 'Threshold and channel interaction in cochlear implant users: Evaluation of the tripolar electrode configuration', *The Journal of the Acoustical Society of America*, 121(3), p. 1642. doi: 10.1121/1.2436712.

Bradley, R. E. (2010) *Predicting Music Enjoyment in Cochlear Implant Users (Doctoral Thesis)*. Washington University School of Medicine.

Brockmeier, S. J. *et al.* (2007) 'Comparison of musical activities of cochlear implant users with different speech-coding strategies.', *Ear and hearing*, 28(2 Suppl), p. 49S–51S. doi: 10.1097/AUD.0b013e3180315468.

Brockmeier SJ, Peterreins M, Lorens A, Vermeire K, Helbig S, Anderson I, Skarzynski H, van de Heyning P, Gstoettner W, Kiefer J (2010) Music perception in electric acoustic stimulation users as assessed by the Mu.S.I.C. Test; in van de Heyning P, Kleine Punte A (eds): *Cochlear Implants and Hearing Preservation*. Adv Otolaryngol. Basel, Karger, vol 67, pp 70–80.

Brockmeier, S. J. *et al.* (2011) 'The MuSIC perception test: A novel battery for testing music perception of cochlear implant users', *Cochlear Implants International*, 12(1), pp. 10–20.

Buechler, M., Lai, W. and Dillier, N. (2004) 'Music perception under bimodal stimulation', *In Conference on Implantable Auditory Prostheses. Pacific Grove, CA; Asilomar.*

Burns, E. M. and Viemeister, N. F. (1981) 'Played-again SAM: Further observations on the pitch of amplitude-modulated noise', *The Journal of the Acoustical Society of America*, 70(6), pp. 1655–1660. doi: 10.1121/1.387220.

Busby, P. a *et al.* (1994) 'Pitch perception for different modes of stimulation using the cochlear multiple-electrode prosthesis.', *The Journal of the Acoustical Society of America*, 95(5 Pt 1), pp. 2658–69. Available at: http://www.ncbi.nlm.nih.gov/pubmed/8207139.

Chen, J. K. *et al.* (2010) 'Music Training Improves Pitch Perception in Prelingually Deafened Children With Cochlear Implants', *Pediatrics*, 125(4), pp. e793–e800. doi: 10.1542/peds.2008-3620.

Cheng, X. *et al.* (2018) 'Music Training Can Improve Music and Speech Perception in Pediatric Mandarin-Speaking Cochlear Implant Users', *Trends in Hearing*, 22, p. 233121651875921. doi:

10.1177/2331216518759214.

de Cheveigné, A. (2005) *Pitch Neural Coding and Perception* Editors: Plack, Christopher J., Oxenham**,** Andrew J., Fay**,** Richard R.

Chorist, M (2005) Rebuilt: How Becoming Part Computer Made Me More Human, Houghton Mifflin Harcourt

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(4), pp. 618–626. doi: 10.1097/AUD.0b013e318174e787.Music.

Cosentino, S. *et al.* (2013) 'Cochlear Implant Filterbank Design and Optimisation: A Simulation Study', *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, (c), pp. 1–1. doi: 10.1109/TASLP.2013.2290502.

Cosentino, S. *et al.* (2016) 'Rate discrimination, gap detection and ranking of temporal pitch in cochlear implant users', *JARO*, 17, pp. 371–382. doi: 10.1007/s10162-016-0569-5.

Crew, J. D., Galvin III, J. J. and Fu, Q.-J. (2012) 'Channel interaction limits melodic pitch perception in simulated cochlear implants.', *Journal of the Acoustical Society of America*, 132(5), pp. EL429-EL435. doi: 10.1121/1.4758770.

Crew, J. D., Galvin, J. J. and Fu, Q.-J. (2015) 'Melodic contour identification and sentence recognition using sung speech', *The Journal of the Acoustical Society of America*, 138(3), pp. EL347-EL351. doi: 10.1121/1.4929800.

Cuddy, L. L. *et al.* (2005) 'Musical difficulties are rare: a study of "tone deafness" among university students.', *Annals of the New York Academy of Sciences*, 1060, pp. 311–24. doi: 10.1196/annals.1360.026.

Cullington, H. E. and Zeng, F.-G. (2010) 'Comparison of bimodal and bilateral cochlear implant users on speech recognition with competing talker, music perception, affective prosody discrimination, and talker identification.', *Ear and Hearing*, 32(1), pp. 16–30. doi: 10.1097/AUD.0b013e3181edfbd2.

Dhanasingh, A. and Jolly, C. (2017) 'An overview of cochlear implant electrode array designs', *Hearing Research*. Elsevier B.V, 356, pp. 93–103. doi: 10.1016/j.heares.2017.10.005.

Donner, A. (1986) A review of inference procedures for the intraclass correlation coefficient in the one way random effects model. *International Statistical Review*, 54(1) pp67-82

Dowling, W. J. and Fujitani, D. S. (1971) 'Contour, interval, and pitch recognition in memory for melodies', *The Journal of the Acoustical Society of America*, 49(2B), pp. 524–531. doi: 10.1121/1.1912382.

Drennan, W. R. *et al.* (2008) 'Discrimination of Schroeder-phase harmonic complexes by normal-hearing and cochlear-implant listeners.', *Journal of the Association for Research in Otolaryngology : JARO*, 9(1), pp. 138–49. doi: 10.1007/s10162-007-0107-6.

Drennan, W. R. *et al.* (2010) 'Sensitivity of psychophysical measures to signal processor modifications in cochlear implant users.', *Hearing research*. Elsevier B.V., 262 Hz(1–2), pp. 1–8. doi: 10.1016/j.heares.2010.02.003.

Drennan, W. R. *et al.* (2015) 'Clinical evaluation of music perception, appraisal and experience in cochlear implant users', *International Journal of Audiology*, 54(2), pp. 114–123. doi: 10.3109/14992027.2014.948219.Clinical.

Drennan, W. R. and Rubinstein, J. T. (2008) 'Music perception in cochlear implant users and its relationship with psychophysical capabilities', *The Journal of Rehabilitation Research and Development*, 45(5), pp. 779–790. doi: 10.1682/JRRD.2007.08.0118.

Fielden, C. A. *et al.* (2015) 'The perception of complex pitch in cochlear implants: A comparison of monopolar and tripolar stimulation', *The Journal of the Acoustical Society of America*, 138(4), pp. 2524–2536. doi: 10.1121/1.4931910.

Filipo, R. *et al.* (2008) 'Music perception in cochlear implant recipients: comparison of findings between HiRes90 and HiRes120.', *Acta oto-laryngologica*, 128(4), pp. 378–81. doi: 10.1080/00016480701796951.

Fitzgerald, H. and Fitzgerald, D. (2006) 'MuSIC Perception Test User Guide'.

Fitzpatrick, R. *et al.* (1998) 'Evaluating patient-based outcome measures for use in clinical trials', *Health Technology Assessment*, 2(14).

Fleiss, J. L. (1999) *The Design and Analysis of Clinical Experiments*. Hoboken, NJ, USA: John Wiley & Sons, Inc. New York. doi: 10.1002/9781118032923.

Foxton, J. M. *et al.* (2004) 'Characterization of deficits in pitch perception underlying "tone deafness"', *Brain*, 127(4), pp. 801–810. doi: 10.1093/brain/awh105.

Galvin, J. J., Fu, Q.-J. and Nogaki, G. (2007) 'Melodic contour identification by cochlear implant listeners', *Ear and Hearing*, 28(3), pp. 302–319. doi: 10.1097/01.aud.0000261689.35445.20.

Galvin, J. J., Fu, Q.-J. and Oba, S. (2008) 'Effect of instrument timbre on melodic contour identification by cochlear implant users.', *The Journal of the Acoustical Society of America*, 124(4), pp. EL189-95. doi: 10.1121/1.2961171.

Galvin, J. J., Fu, Q.-J. and Oba, S. I. (2009) 'Effect of a competing instrument on melodic contour identification by cochlear implant users.', *The Journal of the Acoustical Society of America*, 125(3), pp. EL98-103. doi: 10.1121/1.3062148.

Gantz, B. J., Turner, C. and Gfeller, K. E. (2006) 'Acoustic plus electric speech processing: preliminary results of a multicenter clinical trial of the Iowa/Nucleus Hybrid implant.', *Audiology & neuro-otology*, 11 Suppl 1(suppl 1), pp. 63–8. doi: 10.1159/000095616.

García-Pérez, M. A. (2014) 'Adaptive psychophysical methods for non-monotonic psychometric functions', *Attention, Perception, and Psychophysics*, 76(2), pp. 621–641. doi: 10.3758/s13414-013-0574-2.

Gfeller, K. *et al.* (1997) 'Perception of rhythmic and sequential pitch patterns by normally hearing adults and adult cochlear implant users', *Ear and Hearing*, 18(3), pp. 252–260. Available at: http://journals.lww.com/ear-hearing/Abstract/1997/06000/Perception_of_Rhythmic_and_Sequential_Pitch.8.aspx (Accessed: 10 April 2013).

Gfeller, K. *et al.* (2000) 'Musical backgrounds, listening habits and aesthetic enjoyment of adult cochlear implant recipients', *Journal of the American Academy of Audiology*, 11, pp. 390–406.

Gfeller, K. *et al.* (2002) 'Recognition of familiar melodies by adult cochlear implant recipients and normal-hearing adults', *Cochlear Implants International*, 3(1), pp. 29–53. Available at: http://onlinelibrary.wiley.com/doi/10.1002/cii.50/abstract (Accessed: 10 April 2013).

Gfeller, K. *et al.* (2005) 'Recognition of "real-world" musical excerpts by cochlear implant recipients and normal-hearing adults', *Ear and Hearing*, 26(3), pp. 237–250. Available at: http://www.ncbi.nlm.nih.gov/pubmed/15937406.

Gfeller, K. *et al.* (2008) 'Multivariate predictors of music perception and appraisal by adult cochlear implant users', *Journal of the American Academy of Audiology*, 19(2), pp. 120–134. doi: 10.3766/jaaa.19.2.3.Multivariate.

Gfeller, K. E. and Lansing, C. (1991) 'Melodic, rhythmic, and timbral perception of adult cochlear implant users', *Journal of speech and hearing research*, 34, pp. 916–920. Available at: http://jslhr.highwire.org/cgi/content/abstract/34/4/916 (Accessed: 12 June 2012).

Gfeller, K. E. 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*, 29(1), pp. 18–39. Available at: http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Musical+perception+of+cochlear+implant+users+as+measured+by+the+Primary+Measures+of+Music+Audiation+An+item+analysis#0 (Accessed: 12 June 2012).

Golub, J. S. *et al.* (2012) 'NIH Public Access', *Otology and Neurotology*, 33(2), pp. 147–153. doi: 10.1097/MAO.0b013e318241b6d3.Spectral.

Gordon, E. E. (1979) 'Developmental music aptitude as measured by the primary measures of music audiation', *Psychology of Music*, 7(1), pp. 42–49.

Grasmeder, M. L. (2016) *Optimising Frequency-to-Electrode Allocation for Individual Cochlear Implant Users*. University of Southampton, ISVR, PhD Thesis.

Hedge, C., Powell, G. and Sumner, P. (2017) 'The reliability paradox: Why robust cognitive tasks do not produce reliable individual differences', *Behavior Research Methods*. Behavior Research Methods, pp. 1–21. doi: 10.3758/s13428-017-0935-1.

Henry, J. a and Meikle, M. B. (2000) 'Psychoacoustic measures of tinnitus.', *Journal of the American Academy of Audiology*, 11, pp. 138–155.

Hopyan, T. *et al.* (2001) 'Music skills and the expressive interpretation of music in children with Williams-Beuren syndrome: pitch, rhythm, melodic imagery, phrasing, and musical affect.', *Child neuropsychology : a journal on normal and abnormal development in childhood and adolescence*, 7(1), pp. 42–53. doi: 10.1076/chin.7.1.42.3147.

Huss, M. and Moore, B. C. J. (2005) 'Dead regions and pitch perception', *The Journal of the Acoustical Society of America*, 117(6), pp. 3841–3852. doi: 10.1121/1.1920167.

Irino, T. and Patterson, R. D. (1996) 'Temporal asymmetry in the auditory system', *The Journal of the Acoustical Society of America*, 99(4), pp. 2316–2331. doi: 10.1121/1.423879.

Jacobson, B. D., (2014) *An Analysis of Cochlear Implant Distortion from a User's Perspective* Harvard-MIT Division of Health Sciences and Technology doi: http://dx.doi.org/10.1101/003244

Jeng, F.-C. *et al.* (2011) 'Exponential modeling of human frequency-following responses to voice pitch.', *International journal of audiology*, 50(9), pp. 582–93. doi: 10.3109/14992027.2011.582164.

Jung, K. H. *et al.* (2009) 'Clinical assessment of music perception in Korean cochlear implant listeners.', *Acta Oto-Laryngologica*, 130(6), pp. 716–723. doi: 10.3109/00016480903380521.

Jung, K. H. *et al.* (2012) 'Psychoacoustic Performance and Music and Speech Perception in Prelingually Deafened Children with Cochlear Implants', *Audiology and Neurotology*, 17, pp. 189–197. doi: 10.1159/000336407.

Kaernbach, C. and Demany, L. (1998) 'Psychophysical evidence against the autocorrelation theory of auditory temporal processing', *Journal of the Acoustical Society of America*, 104(4), pp. 2298–2306.

Kang, R. *et al.* (2009) 'Development and validation of the University of Washington Clinical Assessment of Music Perception test', *Ear and Hearing*, 30(4), pp. 411–418. doi: 10.1097/AUD.0b013e3181a61bc0.

Kenway, B. *et al.* (2015) 'Pitch discrimination: An independent factor in cochlear implant performance outcomes', *Otology and Neurotology*, 36(9), pp. 1472–1479. doi: 10.1097/MAO.0000000000000845.

Koch, D. B. *et al.* (2007) 'Using current steering to increase spectral resolution in CII and HiRes 90K users.', *Ear and hearing*, 28(2 Suppl), p. 38S–41S. doi: 10.1097/AUD.0b013e31803150de.

Koelsch, S. *et al.* (2004) 'Music perception in cochlear implant users: an event-related potential study.', *Clinical neurophysiology : official journal of the International Federation of Clinical Neurophysiology*, 115(4), pp. 966–72. doi: 10.1016/j.clinph.2003.11.032.

Kong, Y.-Y. and Carlyon, R. P. (2010) 'Temporal pitch perception at high rates in cochlear implants', *The Journal of the Acoustical Society of America*, 127(5), pp. 3114–3123. doi: 10.1121/1.3372713.

Koo, T. K. and Li, M. Y. (2016) 'A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research', *Journal of Chiropractic Medicine*. Elsevier B.V., 15, pp. 155–163. doi: 10.1016/j.jcm.2016.02.012.

Kuder, G.F., Richardson, M. W., (1937) The theory of the estimation of test reliability *Psychometrika* 2(3) pp 151–160

Lai, W. and Dillier, N. (2002) 'MACarena: a flexible computer-based speech testing environment,

ENT Department, University Hospital, Zurich, Switzerland', in.

Lai, W. and Dillier, N. (2008) 'Investigating the MP3000 coding strategy for music perception', in *Jahrestagung der Deutschen Gesellschaft für Audiologie*, pp. 9–11. Available at: http://www.uzh.ch/orl/dga2008/programm/wissprog/Lai.pdf (Accessed: 9 April 2013).

Lamb, R. (2010) *Test-retest reliability of the South of England Cochlear Implant Centre (SOECIC) Music Test Battery (MTB): an investigation of two response methods*, *Science*. University of Southampton, ISVR, MSc Thesis. Available at: http://eprints.soton.ac.uk/173337/.

Landis, J. R. and Koch, G. G. (1977) 'The Measurement of Observer Agreement for Categorical Data', *Biometrics*, 33(1), p. 159. doi: 10.2307/2529310.

Landsberger, D. M. *et al.* (2014) 'Perceptual changes in place of stimulation with long cochlear implant electrode arrays', *The Journal of the Acoustical Society of America*, 135(2), pp. EL75-EL81. doi: 10.1121/1.4862875.

Landsberger, D. M. and Srinivasan, A. G. (2009) 'Virtual channel discrimination is improved by current focusing in cochlear implant recipients', *Hearing Research*. Elsevier B.V., 254, pp. 34–41. doi: 10.1016/j.heares.2009.04.007.

Lassaletta, L. *et al.* (2007) 'Does music perception have an impact on quality of life following cochlear implantation?', *Acta oto-laryngologica*, 127(7), pp. 682–6. doi: 10.1080/00016480601002112.

Lassaletta, L. *et al.* (2008) 'Changes in listening habits and quality of musical sound after cochlear implantation.', *Otolaryngology--head and neck surgery : official journal of American Academy of Otolaryngology-Head and Neck Surgery*, 138(3), pp. 363–7. doi: 10.1016/j.otohns.2007.11.028.

Law, L. N. C. and Zentner, M. (2012) 'Assessing Musical Abilities Objectively: Construction and Validation of the Profile of Music Perception Skills', *PLoS ONE*, 7(12). doi: 10.1371/journal.pone.0052508.

Levitt, H. (1971) 'Transformed up-□down methods in psychoacoustics', *The Journal of the Acoustical society of America*, 49(2), pp. 467–477. Available at: http://link.aip.org/link/?JASMAN/49/467/1 (Accessed: 10 April 2013).

Li, X. *et al.* (2013) 'Improved perception of music with a harmonic based algorithm for cochlear implants', *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 21(4), pp. 684–694. doi: 10.1109/TNSRE.2013.2257853.

Licklider, J. C. R. (1951) 'A duplex theory of pitch perception.', *Experientia*, 7(4), pp. 128–34. Available at: http://www.ncbi.nlm.nih.gov/pubmed/14831572.

Limb, C. J. *et al.* (2010) 'Auditory Cortical Activity During Cochlear Implant-Mediated Perception of Spoken Language, Melody, and Rhythm', *Journal of the Association for Research in Otolaryngology : JARO*, 143(1), pp. 133–143. doi: 10.1007/s10162-009-0184-9.

Lo, C. Y. (2013) *Melodic contour training and its effect on speech perception for cochlear implant recipients (Master's Thesis)*. Macquarie University, Sydney, Australia.

Loizou, P. C. (1998) 'Mimicking the human ear', *IEEE Signal Processing Magazine*, pp. 101–130.

Loizou, P. C. (2006) 'Speech processing in vocoder-centric cochlear implants.', *Advances in oto-rhino-laryngology*, 64, pp. 109–43. doi: 10.1159/000094648.

Loizou, P. C., Dorman, M. and Tu, Z. (1999) 'On the number of channels needed to understand speech.', *The Journal of the Acoustical Society of America*, 106(4 Pt 1), pp. 2097–103. Available at: http://www.ncbi.nlm.nih.gov/pubmed/10530032.

Looi, V. *et al.* (2008) 'Music perception of cochlear implant users compared with that of hearing aid users.', *Ear and hearing*, 29(3), pp. 421–34. doi: 10.1097/AUD.0b013e31816a0d0b.

Looi, V. and She, J. (2010) 'PERCEPTION OF MUSIC FOR ADULT COCHLEAR IMPLANT USERS : A QUESTIONNAIRE', *International journal of audiology*, 49(2), pp. 116–28. doi: 10.3109/14992020903405987.

Luo, X., Masterson, M. E. and Wu, C.-C. (2014) 'Contour identification with pitch and loudness cues using cochlear implants.', *The Journal of the Acoustical Society of America*, 135(1), pp. EL8-14. doi: 10.1121/1.4832915.

Maarefvand, M., Marozeau, J. and Blamey, P. J. (2013) 'A cochlear implant user with exceptional musical hearing ability.', *International journal of audiology*, 52(6), pp. 424–32. doi: 10.3109/14992027.2012.762606.

Macherey, O. and Carlyon, R. P. (2012) 'Place-pitch manipulations with cochlear implants', *The Journal of the Acoustical Society of America*, 131(3), pp. 2225–2236. doi: 10.1121/1.3677260.Place-pitch.

Macherey, O. and Carlyon, R. P. (2014) 'Cochlear implants', *Current Biology*. Elsevier Ltd, 24(18), pp. R878–R884. doi: 10.1016/j.cub.2014.06.053.

Magnusson, L. (2011) 'Comparison of the fine structure processing (FSP) strategy and the CIS strategy used in the MED-EL cochlear implant system: speech intelligibility and music sound quality.', *International journal of audiology*, 50(4), pp. 279–87. doi: 10.3109/14992027.2010.537378.

Mandell, J., Schulze, K. and Schlaug, G. (2007) 'Congenital amusia: an auditory-motor feedback disorder?', *Restorative neurology and neuroscience*, 25(3–4), pp. 323–34. Available at: http://www.ncbi.nlm.nih.gov/pubmed/17943009.

McDermott, H. J. (2004) 'Music perception with cochlear implants: a review.', *Trends in amplification*, 8(2), pp. 49–82. Available at: http://www.ncbi.nlm.nih.gov/pubmed/15497033.

McDermott, H. J. and McKay, C. M. (1997) 'Musical pitch perception with electrical stimulation of the cochlea.', *The Journal of the Acoustical Society of America*, 101(3), pp. 1622–31. Available at: http://www.ncbi.nlm.nih.gov/pubmed/9069629.

McGraw, K. O. and Wong, S. P. (1996) 'Forming inferences about some intraclass correlation coefficients.', *Psychological Methods*, 1(1), pp. 30–46. doi: 10.1037//1082-989X.1.1.30.

McKay, C. M., McDermott, H. J., Clark, G. M. (1994) ' Pitch percepts associated with amplitude-modulated current pulse trains in cochlear implantees.', *The Journal of the Acoustical Society of America*, 96 (5), Pt. 1, pp. 2664-2673

Mehra, M. (2012) *Which CI Strategy is the Most Successful for Music Appreciation*.

Micheyl, C. *et al.* (2006) 'Influence of musical and psychoacoustical training on pitch discrimination.', *Hearing research*, 219(1–2), pp. 36–47. doi: 10.1016/j.heares.2006.05.004.

Migirov, L., Kronenberg, J. and Henkin, Y. (2009) 'Self-reported listening habits and enjoyment of music among adult cochlear implant recipients.', *The Annals of otology, rhinology, and laryngology*, 118(5), pp. 350–5. Available at: http://www.ncbi.nlm.nih.gov/pubmed/19548384.

Mirza, S. *et al.* (2003) 'Appreciation of music in adult patients with cochlear implants: a patient questionnaire.', *Cochlear implants international*, 4(2), pp. 85–95. doi: 10.1002/cii.68.

Moore, B. C. J. (2012) *An introduction to the psychology of hearing*. Sixth Edition. Bingley Emerald

Moore, B. C. J. (2007). *Cochlear Hearing Loss: Physiological, Psychological and Technical Issues*. Second edition. New Jersey, USA: Wiley.

Muller, R., and Buttner, P. (1994) A critical discussion of intraclass correlation coefficients. *Statistics in medicine* 13, pp2465-2476

National Institute for Health and Care Excellence (2009) 'Cochlear implants for children and adults with severe to profound deafness', (January 2009), p. 3.

Nie, K., Stickney, G., Zeng, F.G. (2005) Encoding frequency modulation to improve cochlear implant performance in noise. IEEE Transactions on Biomedical Engineering, 52(1)

NIH (2016) *NIDCD Fact Sheet Cochlear Implants*. doi: 10.1016/B978-0-12-416556-4.00005-X.

Nimmons, G. L. *et al.* (2008) 'Clinical Assessment of Music Perception in Cochlear Implant Listeners', *Hearing Research*, 29(2), pp. 149–155.

Nogueira, W. *et al.* (2009) 'Signal Processing Strategies for Cochlear Implants Using Current Steering', *EURASIP Journal on Advances in Signal Processing*, 2009, pp. 1–21. doi: 10.1155/2009/531213.

Olszewski, C. *et al.* (2005) 'Familiar melody recognition by children and adults using cochlear implants and normal hearing children.', *Cochlear implants international*, 6(3), pp. 123–40. doi: 10.1002/cii.5.

Oxenham, A. J. (2008) 'Pitch perception and auditory stream segregation: implications for hearing loss and cochlear implants.', *Trends in amplification*, 12(4), pp. 316–31. doi: 10.1177/1084713808325881.

Parncutt, R. and Cohen, A. J. (1995) 'Identification of microtonal melodies: Effects of scale-step size, serial order, and training', *Perception & Psychophysics*, 57(6), pp. 835–846. doi: 10.3758/BF03206799.

Patterson, R. D. and Irino, T. (1998) 'Modeling temporal asymmetry in the auditory system', *J Acoust Soc Am*, 104(5), pp. 2967–2979. doi: 10.1121/1.423879.

Paynter, K. (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, ISVR, MSc Thesis.

Peretz, I. *et al.* (2013) 'A novel tool for evaluating children's musical abilities across age and culture', *Frontiers in Systems Neuroscience*, 7(July), pp. 1–10. doi: 10.3389/fnsys.2013.00030.

Peretz, I., Champod, A. S. and Hyde, K. (2003) 'Varieties of musical disorders The Montreal Battery of Evaulation of Amusia', *Annals of the New York Academy of Sciences*, 999, pp. 58–75.

Peretz, I., Cummings, S. and Dubé, M.-P. (2007) 'The Genetics of Congenital Amusia (Tone

Deafness): A Family-Aggregation Study', *The American Journal of Human Genetics*, 81(3), pp. 582–588. doi: 10.1086/521337.

Philips, B. *et al.* (2012) 'Characteristics and determinants of music appreciation in adult CI users.', *European archives of oto-rhino-laryngology : official journal of the European Federation of Oto-Rhino-Laryngological Societies (EUFOS) : affiliated with the German Society for Oto-Rhino-Laryngology - Head and Neck Surgery*, 269(3), pp. 813–21. doi: 10.1007/s00405-011-1718-4.

Ping, L. *et al.* (2010) 'Temporal envelope and periodicity cues on musical pitch discrimination with acoustic simulation of cochlear implant', in *2010 International Conference on Audio, Language and Image Processing*. IEEE, pp. 710–714. doi: 10.1109/ICALIP.2010.5685065.

Pinna, G. D. *et al.* (2007) 'Heart rate variability measures: a fresh look at reliability', *Clinical Science*, 113(3), pp. 131–140. doi: 10.1042/CS20070055.

Plack, J. and Oxenham, C. J. (2005) *Pitch Neural Coding and Perception* Editors: Plack**,** Christopher J., Oxenham**,** Andrew J., Fay**,** Richard R.

Prins, N & Kingdom, F. A. A. (2009) Palamedes: Matlab routines for analyzing psychophysical data. http://www.palamedestoolbox.org

Reed, R (2011) Personal communication, from *'Hear the music of a cochlear implant & habilitation masterclass',* Cochlear Technology Centre, Mechelen, Belgium, October 2011

Reed, R. (2016) CI music: seeking perfection, accepting reality. *ENT and audiology news* 25(4) https://www.entandaudiologynews.com/media/5277/ent-so16-reed.pdf

Reiss, L. a J. *et al.* (2007) 'Changes in pitch with a cochlear implant over time.', *Journal of the Association for Research in Otolaryngology : JARO*, 8(2), pp. 241–57. doi: 10.1007/s10162-007-0077-8.

Rosen, S. (1992) 'Temporal Information in Speech: Acoustic, Auditory and Linguistic Aspects', *Philosophical Transactions of the Royal Society B: Biological Sciences*, 336(1278), pp. 367–373. doi: 10.1098/rstb.1992.0070.

Rosen, S., Faulkner, A. and Wilkinson, L. (1999) 'Adaptation by normal listeners to upward spectral shifts of speech: Implications for cochlear implants', *The Journal of the Acoustical Society of America*, 106(6), pp. 3629–3636. doi: 10.1121/1.428215.

Rothery, P. (1979) A nonparametric measure of intraclass correlation. *Biometrika* 66(3), pp629-639

Salvi, R. *et al.* (2017) 'Inner hair cell loss disrupts hearing and cochlear function leading to sensory deprivation and enhanced central auditory gain', *Frontiers in Neuroscience*, 10(JAN), pp. 1–14. doi: 10.3389/fnins.2016.00621.

Santurette, S. and Dau, T. (2007) 'Binaural pitch perception in normal-hearing and hearing-impaired listeners', *Hearing Research*, 223(1–2), pp. 29–47. doi: 10.1016/j.heares.2006.09.013.

Saoji, A. a and Litvak, L. M. (2010) 'Use of "phantom electrode" technique to extend the range of pitches available through a cochlear implant.', *Ear and hearing*, 31(5), pp. 693–701. doi: 10.1097/AUD.0b013e3181e1d15e.

Schouten, J. F., Ritsma, R. J. and Lopes Cardozo, B. (1962) 'Pitch of the residue', *The Journal of the Acoustical Society of America*, 34, pp. 1418–1424.

Schuppert, M. *et al.* (2000) 'Receptive amusia: evidence for cross-hemispheric neural networks underlying music processing strategies.', *Brain : a journal of neurology*, 123 Pt 3, pp. 546–59. Available at: http://www.ncbi.nlm.nih.gov/pubmed/10686177.

Sek, A. and Moore, B. (1995) 'Frequency discrimination as a function of frequency, measured in several ways', *The Journal of the Acoustical Society of America*, 97(4), pp. 2479–2486. Available at: http://link.aip.org/link/?JASMAN/97/2479/1 (Accessed: 9 April 2013).

Semal, C. and Demany, L. (2006) 'Individual differences in the sensitivity to pitch direction', *The Journal of the Acoustical Society of America*, 120(6), p. 3907. doi: 10.1121/1.2357708.

Shapiro, S. S. and Wilk, M. B. (1965) 'An Analysis of Variance Test for Normality (Complete Samples)', *Biometrika*, 52(3/4), p. 591. doi: 10.2307/2333709.

Shrout, P. E. and Fleiss, J. L. (1979) 'Intraclass correlations: uses in assessing rater reliability', *Psychological Bulletin*, 86(2), pp. 420–428.

Smith, D. W. *et al.* (1987) 'Effects of Outer Hair Cell Loss on the Frequency-Selectivity of the Patas Monkey Auditory-System', *Hearing Research*, 29, pp. 125–138.

Smith, Z., M., Delgutte, B., and Oxenholm, A., J., Chimaeric sounds reveal dichotomies in auditory perception (2002), Nature 416, p87-90

Spitzer, J. B., Mancuso, D. and Cheng, M.-Y. (2008) 'Development of a Clinical Test of Musical Perception: Appreciation of Music in Cochlear Implantees (AMICI)', *Journal of the American Academy of Audiology*, 19(1), pp. 56–81. doi: 10.3766/jaaa.19.1.6.

Stainsby, T. H. *et al.* (1997) 'Preliminary Results on Spectral Shape Perception and Discrimination of Musical Sounds by Normal Hearing Subjects and Cochlear Implantees', in Zeng, F.-G. (ed.) *Proceedings of the International Computer Music Conference*, pp. 2–5.

Stamou, L., Schmidt, C. P. and Humphreys, J. T. (2010) 'Standardization of the Gordon Primary Measures of Music Audiation in Greece', *Journal of Research in Music Education*, 58(1), pp. 75–89. doi: 10.1177/0022429409360574.

www.stattrek.com http://stattrek.com/online-calculator/binomial.aspx) accessed May 2018

Stevens, S. S. (1935) 'The Relation of Pitch to Intensity', *The Journal of the Acoustical Society of America*, 150(1935).

Stickney, G. *et al.* (2004) 'Temporal fine structure: the missing component in speech processing algorithms', *International Congress Series*, 1273, pp. 23–26. doi: 10.1016/j.ics.2004.09.017.

Sucher, C. M. and Mcdermott, H. J. (2009) 'Bimodal stimulation : benefits for music perception and sound quality', *Cochlear Implants International*, 10(S1), pp. 96–99. doi: 10.1002/cii.

Svirsky, M. (2017) 'Cochlear implants and electronic hearing', *Physics Today*, 70(8), pp. 53–58. doi: 10.1063/PT.3.3661.

Swanson, B. A. (2008) *Pitch Perception with Cochlear Implants (Doctoral Thesis)*. The University of Melbourne.

Taber, K. S. (2017) 'The Use of Cronbach's Alpha When Developing and Reporting Research Instruments in Science Education', *Research in Science Education*. Research in Science Education, pp. 1–24. doi: 10.1007/s11165-016-9602-2.

Townshend, B. *et al.* (1987) 'Pitch perception by cochlear implant subjects.', *The Journal of the Acoustical Society of America*, 82(1), pp. 106–15. Available at: http://www.ncbi.nlm.nih.gov/pubmed/3624633.

Ukoumunne, O.C., (2002) A comparison of confidence interval methods for the intraclass correlation coefficient in cluster randomized trials. *Statistics in Medicine, 21, pp* 3757–3774 (DOI: 10.1002/sim.1330)

Vandali, A. E. *et al.* (2005) 'Pitch ranking ability of cochlear implant recipients: A comparison of sound-processing strategies', *The Journal of the Acoustical Society of America*, 117(5), p. 3126. doi: 10.1121/1.1874632.

Vaz, S. *et al.* (2013) 'The Case for Using the Repeatability Coefficient When Calculating Test-Retest Reliability', *PLoS ONE*, 8(9), pp. 1–7. doi: 10.1371/journal.pone.0073990.

Veekmans, K. *et al.* (2009) 'Comparison of music perception in bilateral and unilateral cochlear implant users and normal-hearing subjects.', *Audiology & neuro-otology*, 14(5), pp. 315–26. doi:

10.1159/000212111.

Verschuure, J. and Van Meeteren, a a (1975) 'The effect of intensity on pitch', *Acustica*, 32(1), pp. 33–44.

Wang, W., Zhou, N. and Xu, L. (2011) 'Musical pitch and lexical tone perception with cochlear implants.', *International journal of audiology*, 50(4), pp. 270–8. doi: 10.3109/14992027.2010.542490.

Weibull.com http://www.weibull.com/hotwire/issue56/relbasics56.htm accessed May 2018

Weibull.com http://www.weibull.com/hotwire/issue14/relbasics14.htm accessed May 2018

Wichmann, F. a and Hill, N. J. (2001) 'The psychometric function: I. Fitting, sampling, and goodness of fit.', *Perception & psychophysics*, 63(8), pp. 1293–313. Available at: http://www.ncbi.nlm.nih.gov/pubmed/11800458.

Wilson, B.S., Sun, X., Schatzer, R.,Wolford, R.D. (2004) 'Representation of fine structure or fine frequency information with cochlear implants', *International Congress Series* 1273, pp3 – 6

Won, J. H. *et al.* (2010) 'Psychoacoustic abilities associated with music perception in cochlear implant users.', *Ear and hearing*, 31(6), pp. 796–805. doi: 10.1097/AUD.0b013e3181e8b7bd.

Won, J. H. *et al.* (2015) 'Spectral and Temporal Analysis of Simulated Dead Regions in Cochlear Implants', *JARO - Journal of the Association for Research in Otolaryngology*, 16(2), pp. 285–307. doi: 10.1007/s10162-014-0502-8.

Wong, L. L. N. *et al.* (2008) 'New cochlear implant coding strategy for tonal language speakers.', *International journal of audiology*, 47(6), pp. 337–47. doi: 10.1080/14992020802070788.

Wright, R. and Uchanski, R. M. (2012) 'Music perception and appraisal: cochlear implant users and simulated cochlear implant listening.', *Journal of the American Academy of Audiology*, 23(5), p. 350–65; quiz 379. doi: 10.3766/jaaa.23.5.6.

Xu, L., and Pfingst, B. E. (2003) 'Relative importance of temporal envelope and fine structure in lexical-tone perception (L).', *Journal of the Acoustical Society of America,* 114 (6), Pt. 1., pp. 3024-3027

Yitao, M. and Li, X. (2013) 'Music and Cochlear Implants', *Journal of Otology*, 8(1), pp. 32–38. doi: 10.1016/S1672-2930(13)50004-3.

Yost, W. A. (2009) 'Pitch perception', *Attention, Perception, & Psychophysics*, 71(8), pp. 1701–1715. doi: 10.3758/APP.

Yukawa, K. *et al.* (2004) 'Effects of insertion depth of cochlear implant electrodes upon speech perception.', *Audiology & neuro-otology*, 9(3), pp. 163–72. doi: 10.1159/000077267.

Zarate, J. M., Ritson, C. R. and Poeppel, D. (2013) 'The Effect of Instrumental Timbre on Interval Discrimination', *PLoS ONE*, 8(9). doi: 10.1371/journal.pone.0075410.

Zeng, F.-G. *et al.* (2005) 'Speech recognition with amplitude and frequency modulations', *Proceedings of the National Academy of Sciences*, 102(7), pp. 2293–2298. doi: 10.1073/pnas.0406460102.

Zeng, F.-G., Tang, Q. and Lu, T. (2014) 'Abnormal pitch perception produced by cochlear implant stimulation.', *PloS one*, 9(2), p. e88662. doi: 10.1371/journal.pone.0088662.

Zhang, F., Benson, C. and Cahn, S. J. (2013) 'Cortical encoding of timbre changes in cochlear implant users.', *Journal of the American Academy of Audiology*, 24(1), pp. 46–58. doi: 10.3766/jaaa.24.1.6.

Zhou, H., Muellerleile, P., Ingram, D., Wong, S.P. (2011) Confidence Intervals and F Tests for Intraclass Correlation Coefficients Based on Three-Way Mixed Effects Models *Journal of Educational and Behavioural Statistics* 36(5) pp. 638-671