

CLINICAL APPLICABILITY
OF ADAPTIVE SPEECH TESTING:
A COMPARISON OF THE ADMINISTRATION TIME,
ACCURACY, EFFICIENCY AND RELIABILITY
OF ADAPTIVE SPEECH TESTS
WITH CONVENTIONAL
SPEECH AUDIOMETRY

A thesis submitted in partial fulfilment of the requirements for the

Degree

of Master of Audiology

in the University of Canterbury

by B. P. Sincock

University of Canterbury

2008

Abstract

Adaptive procedures are a common method of investigating sensory abilities in research settings; however, their use in clinical settings is more limited. Little research has been done investigating the implementation of adaptive procedures into Audiological speech tests, and to date, no studies have compared and evaluated adaptive speech tests with current clinical speech audiometry. This study investigated the advantages of using both closed-set and open-set adaptive speech tests in the clinical Audiology setting, with respect to administration time, accuracy, efficiency and reliability. Preliminary testing of the two major adaptive procedures (staircase and maximum-likelihood procedures) was conducted using a selection of different parameters chosen on the basis of the results of previous research (Kaernbach, 1991; García-Pérez, 1998) to determine the optimal procedures and parameters for use in clinical speech tests. Focus was given to the staircase procedures, with comparisons made between tests using variations in step size – constant step sizes and larger step sizes at the beginning – and different termination criteria. It was found that both adaptive closed-set staircase tests (with both step size variations) performed similarly, whereas the adaptive open-set staircase test with larger step sizes at the beginning showed advantages over the equivalent constant step size test in terms of administration time, accuracy and efficiency. The maximum-likelihood QUEST procedure showed advantages over the staircase procedures in terms of administration time; however, the reliability of both this test and conventional speech audiometry was poor, indicating that these tests are not the most suitable tests for a clinical setting. Subsequent clinical testing of the *optimal* adaptive speech tests using participants with varying degrees of hearing loss found that administration time was similar between conventional speech audiometry and the adaptive closed-set staircase tests when the optimal termination criteria identified in the Preliminary Testing Phase were employed. The adaptive open-set staircase test with larger step sizes at the beginning showed the best accuracy of any of the tests when using the pure-tone average as a reference, while the efficiency of all the adaptive staircase tests was similar. Overall, the results highlight some of the potential advantages of adaptive speech testing in the clinical Audiology setting; however, further studies are required to determine the specific parameters that achieve the best results.

Acknowledgements

I would like to thank my primary supervisor Dr. Greg O’Beirne for his guidance and support with this thesis. His work designing and adapting the University of Canterbury Monosyllabic Adaptive Speech Test (UC MAST) program that formed the basis of this thesis was very much appreciated.

I also wish to thank my secondary Supervisor Dr. Emily Lin for helping with the statistical analyses of the results of this study. I greatly appreciate her patience and the knowledge she contributed.

I wish to acknowledge the participants who willingly gave their time to participate in this study. Their time and co-operation was much appreciated.

This thesis is dedicated to my husband Jared and my 6-week old baby son Asher. Without my husband’s love, support and patience, this project would not have been possible. I thank him especially for his help with the formatting of this thesis, and the incredible automatic back-up program which he designed and put in place to ensure that the many computer problems I experienced did not lead to any major data losses. I also thank my baby Asher for providing me with the motivation to finish this work (and for sleeping at least part of each day so I had time to write it up)!

Table of Contents

1. Introduction	3
1.1 The Importance of Speech Audiometry	4
1.2 Speech Audiometry Response Formats: Open-Set versus Closed-Set Speech Tests.....	4
1.3. Non-Adaptive Procedures: <i>The Method of Constant Stimuli</i>	5
1.4 Adaptive Procedures	7
1.4.1 History of Adaptive Procedures in Audiology.....	7
1.4.2 Categories of Adaptive Procedures	8
1.4.2.1 Staircase Procedures	8
1.4.2.2 Maximum-Likelihood Procedures	11
1.5 Comparing Procedures.....	13
1.5.1 Comparison between Adaptive Methods and the Method of Constant Stimuli	14
1.5.1.1 Administration Time of Adaptive Procedures versus Methods of Constant Stimuli	14
1.5.1.2 Accuracy of Adaptive Procedures versus Methods of Constant Stimuli	15
1.5.1.3 Efficiency of Adaptive Procedures versus Methods of Constant Stimuli	16
1.5.1.4 Reliability of Adaptive Procedures versus Methods of Constant Stimuli	17
1.5.2 Comparison between Staircase Procedures and Maximum-Likelihood Procedures	18
1.5.2.1 Administration Time of Staircase Procedures versus Maximum-Likelihood Procedures	18
1.5.2.2 Accuracy of Staircase Procedures versus Maximum-Likelihood Procedures.....	18
1.5.2.3 Efficiency of Staircase Procedures versus Maximum-Likelihood Procedures.....	19
1.5.2.4 Reliability of Staircase Procedures versus Maximum-Likelihood Procedures.....	19

1.6 Optimal Design Criteria and Parameters in Adaptive Procedures	20
1.6.1 Open-Set versus Closed-Set Response Formats	20
1.6.2 Phonemic Scoring versus Whole-Word Scoring.....	21
1.6.3 Starting Level.....	21
1.6.4 Target Levels.....	22
1.6.5 Termination Criteria.....	22
1.6.6 Number of Trials.....	22
1.6.7 Parameters Specifically Relating to Staircase Procedures.....	24
1.6.8 Parameters Specifically Relating to Maximum-Likelihood Procedures.....	25
1.7 Adaptive Speech Tests	26
1.8 Aims of this Study	26
1.9 Hypotheses.....	27
1.10 Structure of the Study and Thesis Organization.....	28
2. Selection of Test Materials, Procedures and Parameters for the Preliminary Testing of the Adaptive Tests	33
2.1 Test Materials - Closed-Set and Open-Set Speech Tests	33
2.2 Selection of Adaptive Procedures for Implementation into the Closed-Set Speech Tests	34
2.2.1 Staircase Procedure	34
2.2.2 Maximum-Likelihood Procedure	35
2.3 Selection of Adaptive Procedures for Implementation into the Open-Set Speech Tests	36
2.4 Selection of Other Parameters for use in the Adaptive Staircase Procedures	37
2.4.1 Starting Level.....	37
2.4.2 Step Size - Constant versus Larger at the Beginning	37
2.4.3 Initial and Working Step Sizes - Absolute Values.....	38
2.4.4 Termination Criteria.....	38
2.4.5 Practice Reversals and Real Reversals - Threshold Calculation	40
2.5 Selection of General Parameters for use in the Adaptive Maximum-Likelihood Procedure - QUEST.....	42

3. General Methods.....	47
3.1 Participants	47
3.2 Equipment	48
3.2.1 UC MAST Software	49
3.2.2 Calibration of Equipment	51
3.3 Test Procedure.....	52
3.4 Measures.....	55
3.4.1 Administration Time.....	55
3.4.2 Accuracy	56
3.4.3 Efficiency.....	56
3.4.4 Reliability	57
3.5 Data Analysis	57
4. Preliminary Testing Phase for the Refinement of the Adaptive Procedures.....	61
4.1 Introduction	61
4.2 Methods	61
4.2.1 Participants	61
4.2.2 Speech Tests	61
4.3 Results	62
4.3.1 Administration Time	62
4.3.1.1 Administration of Instructions	63
4.3.1.2 Administration of Testing Procedure - Staircase Speech Tests	64
4.3.1.3 Administration of Testing Procedure - Maximum-Likelihood (QUEST) Speech Tests	76
4.3.2 Accuracy	79
4.3.3 Efficiency - Staircase Tests.....	83
4.3.4 Reliability - Staircase and Maximum-Likelihood (QUEST) Tests	85
4.4 QUEST Slope Estimates - 4AFC Tasks.....	87
4.5 Summary of Main Findings	87
4.5.1 Closed-Set Staircase Procedures.....	88
4.5.2 Open-Set Staircase Procedures	89

4.5.3 Maximum-Likelihood Procedure/Estimates	90
4.6 Conclusions - Speech Tests Chosen for Clinical Testing Phase	91
5. Clinical Testing Phase of the Optimal Adaptive Procedures	95
5.1 Methods.....	95
5.1.1 Participants.....	95
5.1.2 Speech Tests.....	95
5.1.3 Special Testing Circumstances.....	96
5.2 Results	97
5.2.1 Administration Time - Testing Procedures.....	97
5.2.2 Accuracy	98
5.2.3 Efficiency	101
5.3 Summary of Main Findings.....	103
6. General Discussion.....	107
6.1 Comparison between Adaptive Methods and the Method of Constant Stimuli	107
6.1.1 <i>Administration Time of Adaptive Procedures versus Conventional Speech Audiometry</i>	107
6.1.2 <i>Accuracy of Adaptive Procedures versus Conventional Speech Audiometry</i>	108
6.1.3 <i>Efficiency of Adaptive Procedures versus Conventional Speech Audiometry</i>	109
6.1.4 <i>Reliability of Adaptive Procedures versus Conventional Speech Audiometry</i>	109
6.1.5 Additional Advantages of Adaptive Procedures	110
6.1.6 Additional Disadvantages of Adaptive Procedures	111
6.2 Comparison between Adaptive Procedures: Staircase Procedures versus Maximum-Likelihood Procedures	111
6.2.1 <i>Administration Time of Staircase Procedures versus Maximum-Likelihood Procedures</i>	111
6.2.2 <i>Accuracy of Staircase Procedures versus Maximum-Likelihood Procedures</i>	112

6.2.3 <i>Efficiency of Staircase Procedures versus Maximum-Likelihood Procedures</i>	113
6.2.4 <i>Reliability of Staircase Procedures versus Maximum-Likelihood Procedures</i>	114
6.3 Comparison between Closed-Set and Open-Set Adaptive Speech Tests	114
6.4 Comparison between Step Size Variations in the Staircase Procedures: Constant versus Larger at the Beginning	118
6.5 Implications for Clinical Speech Audiometry	119
6.6 Limitations of the Study	119
6.6.1 <i>Validity of Accuracy Comparisons</i>	120
6.6.2 <i>Equipment Limitations</i>	121
6.6.3 <i>UC MAST Program Limitations</i>	122
6.7 Directions for Future Research	123
6.8 Conclusions	124
7. Appendices	129
Appendix I: Project Information Sheets, Consent Forms and Participant Questionnaire.....	129
Appendix II: Instructions to Participants	135
Appendix III: Example of a Complete Test Data Set.....	137
8. References	143

Chapter 1
Introduction

1. Introduction

Adaptive procedures have been shown to be efficient, accurate and reliable in the determination of thresholds when utilized in research settings (Linschoten, Harvey, Eller, & Jafek, 2001; Zera, 2004); however, there are few studies supporting the usefulness of these procedures in clinical settings. If the advantages of adaptive procedures (in terms of administration time, accuracy, efficiency and reliability) can be incorporated into an Audiological, adaptive speech test, it would undoubtedly be beneficial in clinical practice.

Current practice in New Zealand Audiology clinics is to use a method of constant stimuli to administer speech tests, whereby lists of words at fixed intensities are presented to participants and the percentage of correctly identified phonemes is obtained. Adaptive procedures are those in which the presentation of each stimulus is based on the previous one or more stimuli and the participant's responses to them. These adaptive procedures vary the intensity of the input stimuli to determine the level at which a predetermined percentage correct score occurs. The current study compared the usefulness of two forms of adaptive speech tests with that of conventional non-adaptive speech audiometry. As well as implementing an adaptive procedure into a forced-choice task, the current study also utilized adaptive procedures by implementing them into an open-set task. This implementation was innovative for two main reasons: (1) The tests had an open-set format, whereas in the past, adaptive procedures were more commonly investigated using closed-set, forced-choice or yes/no tasks; (2) The adaptive procedure was based on a phonemic scoring system (in which each response by the participant was scored according to the *degree* of correctness) which influenced the presentation level of the next stimulus word. In all previous studies (e.g. Bernstein & Gravel, 1990; Levitt, 1971; Mackie & Dermody, 1986), responses had been scored as either 'completely correct' or 'completely incorrect', and these two alternatives alone determined the course that the experiment would take.

1.1 The Importance of Speech Audiometry

Speech audiometry is a routinely performed part of the Audiological test battery. An important aim of clinical speech audiometry is to determine the Speech Reception Threshold (SRT), which is the level at which an individual can recognize 50% of speech sounds. The SRT is an important measure in clinical Audiology as it is considered to be an indication of a person's hearing sensitivity for speech sounds. It provides a measure that can be used diagnostically to examine an individual's speech processing abilities throughout the auditory system. Although the pure-tone average (PTA; typically calculated by averaging the pure-tone air-conduction thresholds at 500 Hz, 1 kHz and 2 kHz) can be used to gauge an individual's hearing sensitivity to speech sounds, this measure is based on thresholds for simple pure-tones, rather than complex speech sounds, and is, therefore, not typical of real-world listening situations. The SRT also provides a means to cross-check the validity of the air-conduction pure-tone audiometry results, making it an early and powerful indicator of test consistency (Olsen & Matkin, 1991). The reliability of behavioural data is inferred from the degree of consistency between the SRT value and PTA. In practice, if the SRT is within ± 6 dB of the PTA it is considered to be in good agreement; between ± 7 dB and ± 12 dB fair agreement; and greater than ± 13 dB poor agreement (Brandy, 2002). Disagreement between the SRT and the PTA can be an indication of inconsistencies in the test results. According to the American Speech-Language-Hearing Association [ASHA] Committee on Audiologic Evaluation (1988), such inconsistencies whereby the SRT is lower than the PTA may be due to test variables such as equipment malfunction or misunderstanding of instructions by the participant. Additionally, a SRT that is significantly lower than the PTA can be an indication of pseudohypacusis (Olsen & Matkin, 1991), while a SRT significantly higher than the PTA can be an indication of a retrocochlear lesion or some other central auditory disorder (Crandell, 1991), or a language disorder (Silman & Silverman, 1997).

1.2 Speech Audiometry Response Formats: Open-Set versus Closed-Set Speech Tests

There are two different response formats that can be utilized in speech tests – open-set and closed-set. Open-set tests are those in which a participant listens to a target word or sentence and repeats it back. Closed-set tests, also known as forced-choice tests, are multiple-choice

tasks, in which a participant listens to a target word and chooses a response from a set of pre-assigned responses, usually in written or pictorial form.

The open-set response format, which is used for the conventional Audiological speech test in New Zealand, is advantageous, as the results can be seen as more representative of real-world listening performance. The open-set response format is somewhat flexible in the type of scoring method used, as responses can be scored according to whole words correct or phonemes correct (as is the case in the conventional Audiological speech test used in New Zealand).

A number of closed-set speech recognition tests have been developed over the years, most commonly for use with children. These include Word Intelligibility by Picture Identification (Lerman, Ross, & McLauchlin, 1965; Ross & Lerman, 1970), the California Consonant Test (Owens & Schubert, 1977), the University of Oklahoma Closed-Response Speech Test (Pederson & Studebaker, 1972), and Northwestern University Children's Perception of Speech (Katz & Elliot, 1978). Closed-set tests usually comprise a word or picture pointing task, and therefore the accuracy of results is not influenced by the participant's expressive vocabulary, articulation or the intelligibility of their voice. Additionally, closed-set tests reduce learning effects when repeated measures are required (Gelfand, 2001). A disadvantage of closed-set tests, however, is that the participant's response is limited to one of the pre-assigned responses. Even if the individual does not hear any of the given words, he/she is forced to choose a response from the given alternatives, thus introducing an element of chance to the procedure, the size of which is dependent on the number of alternative responses available.

1.3 Non-Adaptive Procedures: *The Method of Constant Stimuli*

The method of constant stimuli is a non-adaptive method, in which the distribution of trials (e.g. words in the case of speech testing) at different intensities is specified before testing begins. The majority of speech recognition tests are carried out using methods of constant stimuli, as lists of words are presented at a number of fixed intensities, and percentage correct scores are obtained.

The conventional speech test for adults most commonly used by New Zealand Audiologists is Boothroyd and Nittouer's meaningful consonant-vowel-consonant (CVC) words (1988). This is an open-set test which employs the method of constant stimuli, and requires participants to verbally repeat target monosyllabic words presented in lists of ten. In New Zealand it is the standard practice to present the lists of words monaurally at three different intensities, in order to obtain three percentage correct scores, which provide a good estimate of the patient's performance-intensity curve (psychometric function). As shown in Figure 1, this performance-intensity speech curve plots intensity on the abscissa and the proportion of correct responses on the ordinate.

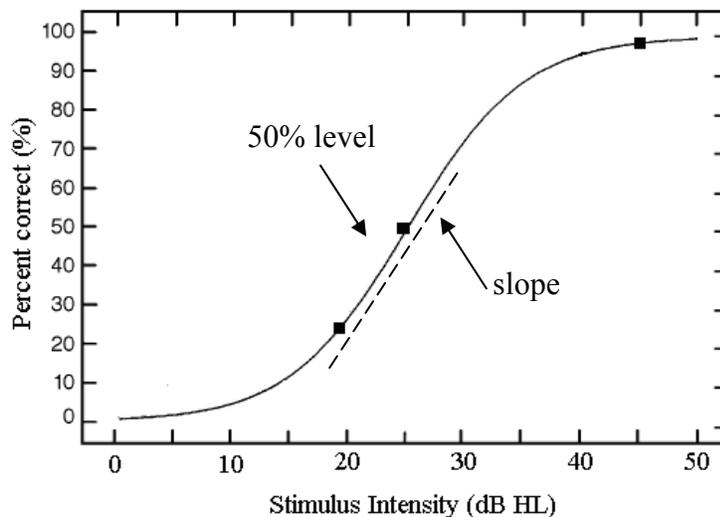


Figure 1. Example of a performance-intensity speech curve/psychometric function for one ear (Adapted from Leek, 2001), plotting percentage of phonemes correct as a function of stimulus intensity. As shown, three points are generally determined in order to approximate the curve.

Conventional speech audiometry, yielding a three-point psychometric function for each ear, requires participants to listen to and repeat back (one at a time) a total of at least 60 words. These words are presented via a compact disc at a constant rate, and therefore the participant is required to answer within a given time frame (although the examiner is able to pause the C.D. to allow more time). The test also requires the participant to have adequate verbal language skills, and relies on the intelligibility of their speech, as well as their perception of speech, to obtain accurate results. Responses are scored according to the number of phonemes repeated correctly; thus each percentage correct score is the result of 30 scored items (10 words/3 phonemes each).

1.4 Adaptive Procedures

Adaptive procedures form a class of sequential experiments, which rely on the experimental data gathered during the test to determine the course that the experiment will take. Unlike non-adaptive methods utilizing stimuli intensities which are fixed prior to testing, adaptive procedures are those in which a component of the stimulus (e.g. level) on any one trial is determined by one or more of the preceding stimuli and the participant's responses (Falmagne, 1986). In the case of adaptive speech testing, an adaptive procedure is used to find the stimulus intensity at which the participant's response is correct a specified percentage of the time.

Adaptive procedures are efficient at targeting the reception or detection threshold of an individual, because trials are placed at a targeted point on a psychometric function, which does not have to be estimated before testing begins. In the case of adaptive speech testing, threshold is defined as the signal level at which the probability of a correct response is halfway between perfect performance (100%) and chance performance (Kaernbach, 2001). In an open-set test, the threshold value will be 50%, as chance performance can be as low as 0%. In a closed-set test, however, chance performance will depend on the number of alternative responses that are available, and therefore, the threshold value will be higher than 50%. Adaptive procedures cited in the literature have only been applied to forced-choice tasks, in particular two alternative forced-choice (2AFC) tasks and yes/no tasks.

1.4.1 History of Adaptive Procedures in Audiology

Adaptive procedures have been utilized in clinical Audiology since the field began. An adaptive 'staircase' procedure was the first adaptive procedure to have a clinical application in Audiology, when it was employed by Hughson and Westlake (1944) to determine auditory thresholds for pure-tones. A modified version of this original method (Carhart & Jerger, 1959) is still used in clinics today. The method involves using a down-10 dB and up-5 dB approach. With every correct response (i.e. accurate detection of a tone), the stimulus intensity is lowered 10 dB until the participant no longer responds, after which the intensity is increased in 5 dB steps until a response is made. As the intensity of each pure-tone is

determined by the participant's response to the preceding stimuli, this test is classed as an *adaptive* procedure.

The American Speech-Language-Hearing Association [ASHA] Committee on Audiologic Evaluation (1988) also recommends the use of a rather simplistic adaptive procedure for the determination of an individual's SRT, although this procedure is not commonly used in New Zealand Audiology clinics. During the preliminary phase of the procedure, for the determination of an appropriate starting level, spondaic words (which the participant is familiarized with prior to testing) are presented at 10 dB decrements until two words at the same intensity are missed. If only one word is missed, another spondee is presented at the same level. Although simple, this procedure is essentially adaptive, because the level of presentation of each word is dependent on whether the individual responded correctly to the previous word.

1.4.2 Categories of Adaptive Procedures

There are two main categories of adaptive procedures that are commonly employed in current psychophysical testing procedures - staircase procedures and maximum-likelihood procedures. Brief mention will also be made to a third category of adaptive procedures - the Parameter Estimation by Sequential Testing (PEST) procedure - which preceded the emergence of maximum-likelihood procedures. While the PEST procedure is considered by many authors to be in a category of its own (Leek, 2001; Shelton, Picardi, & Green, 1982), it will not be discussed at length here, because of the fact that it has been largely superseded by more efficient procedures such as QUEST (Watson & Pelli, 1983).

1.4.2.1 Staircase Procedures

As mentioned above, simple staircase procedures (Dixon & Mood, 1948) that target the 50% correct level on a psychometric function, involve the reduction of stimulus intensity after every correct response and the increase of stimulus intensity after every incorrect response. The amount by which the level of the stimulus is increased or decreased is referred to as a *step*, and in the case of the simple staircase procedure, the step size is identical for

increases and decreases in intensity (Figure 2). A series of steps in either the positive or negative direction is referred to as a *run*.

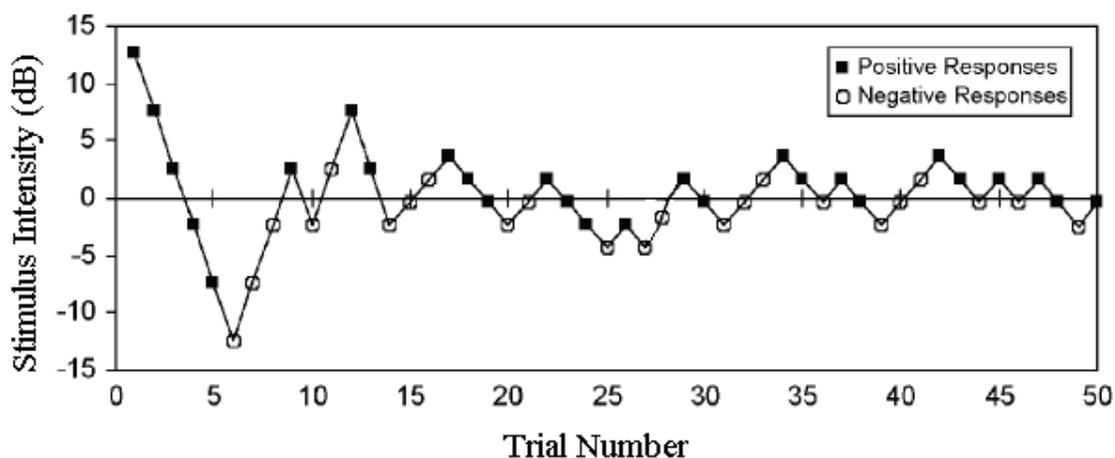


Figure 2. A simple staircase adaptive track (Adapted from Leek, 2001). Decibel values are relative to an arbitrary threshold of 0 dB, which is represented by the horizontal line.

An estimate of the individual's threshold can then be calculated, typically by averaging the levels at the reversals (points where the runs change direction). Different researchers suggest using different numbers of reversals to obtain a threshold estimate. Wetherill and Levitt (1965), for example, recommend continuing the adaptive test until at least six or eight reversals are obtained; however, some studies have used as few as four reversals (Bernstein & Gravel, 1990; Nagy & Kamholz, 1995; Scheffrin, Shinomori, & Werner, 1995; Zanker & Hüpogens, 1994), and others as many as forty reversals (Heidenrich & Turano, 1996). In many cases, the accuracy and reliability of the threshold estimate may be improved (at the expense of time) by obtaining a larger number of reversals (Bernstein & Gravel, 1990).

The simple staircase procedure is useful for determining the 50% performance level on a psychometric function; however, more points on the curve are often required to define its parameters and estimate its slope. Levitt (1971) reviewed a transformation of the simple adaptive staircase procedure, first introduced by Zwislocki, Maire, Feldman and Rubin (1958) and later popularized by Wetherill and Levitt (1965), that would enable specific locations on a psychometric function to be targeted. Unlike the simple staircase procedure that resulted in a change in stimulus intensity after every response, Levitt's 'transformed' procedure used *sequences* of responses to initiate a change in stimulus intensity. To converge

on a percentage correct value higher than 50% on the psychometric curve, a sequence of correct responses must be obtained before stimulus intensity is decreased; however, a single incorrect response results in an increase in stimulus intensity. An example of this transformed adaptive staircase procedure is the three-down, one-up procedure (shown in Figure 3), in which three correct answers are required to decrease stimulus intensity, while only one incorrect answer is required to increase stimulus intensity. This three-down, one-up procedure utilizes fixed step sizes and targets the 79.4% correct level.

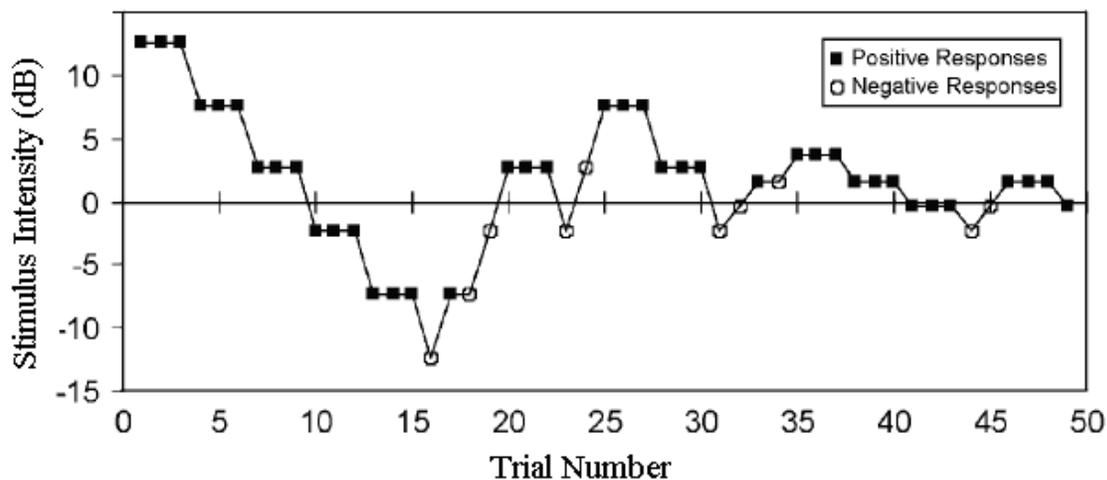


Figure 3. Transformed staircase adaptive track, following the three-down, one-up rule (Adapted from Leek, 2001). Decibel values are relative to an arbitrary threshold of 0 dB, which is represented by the horizontal line.

The transformed adaptive staircase procedure has been implemented into a number of systems that have been used in research. Bernstein and Gravel (1990) demonstrated that their computer-assisted Interweaving Staircase Procedure which utilized a two-down, one-up rule targeting the 70.7% correct level gave sufficiently accurate pure-tone threshold estimates (calculated from four reversals) when tested with adults, after an average of 43.3 trials ($S.D. = 8.0$).

Kaernbach (1991) described a method for determining a greater number of values on a psychometric curve than was possible using Levitt's (1971) transformed staircase procedures. Kaernbach's method, known as the weighted up/down staircase procedure, involves the use of different sized steps for increases and decreases in intensity. Unlike transformed staircase

procedures, which can only target performance levels that are able to be estimated by specific sequences of up and down trials, weighted staircase procedures can target any performance level. Additionally, Kaernbach (1991) reported a 10% time saving when using a weighted up/down staircase procedure in place of a transformed staircase procedure.

1.4.2.2 Maximum-Likelihood Procedures

Maximum-likelihood procedures arose from an earlier form of adaptive procedure known as PEST (Taylor & Creelman, 1967). The PEST procedure utilizes an algorithm that changes step sizes and track direction in order to focus an adaptive track towards its target threshold (Figure 4). The threshold estimate is simply the final value determined by the trial placement procedure once the estimate has been adequately defined.

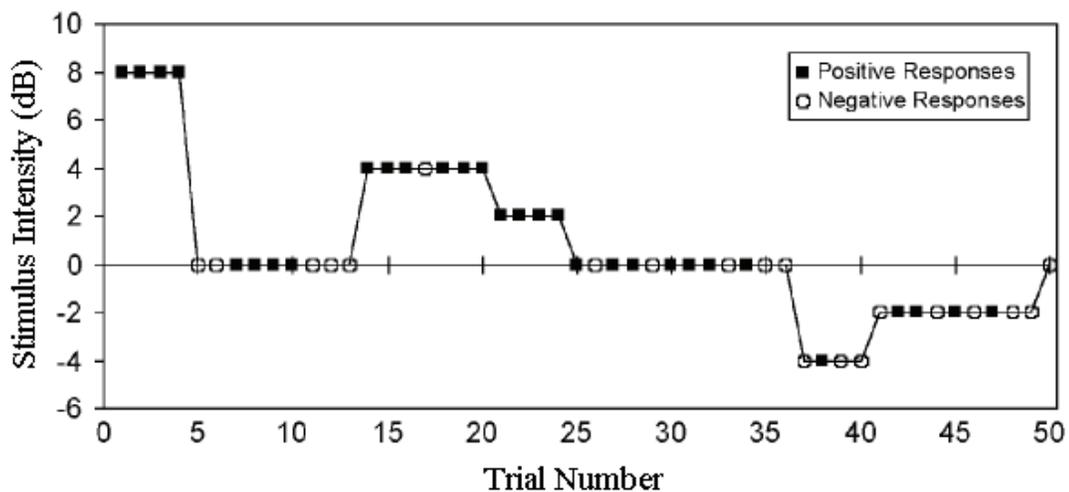


Figure 4. PEST adaptive track (Adapted from Leek, 2001). Decibel values are relative to an arbitrary threshold of 0 dB, which is represented by the horizontal line.

A number of modifications to the original PEST procedure were made over the years, in an attempt to improve efficiency (Hall, 1981; Pentland, 1980). Hall (1981) proposed a hybrid procedure, which utilized the PEST rules for trial placement, but rather than simply taking the final trial placement value as threshold, used data from all the trials to construct a psychometric curve from which a threshold could be extracted. This hybrid procedure had advantages insofar as it was not as susceptible to participants' lapses in concentration or the values chosen for the step size and starting value, and was able to provide a slope estimate of the psychometric curve as well as the threshold estimate.

From further modifications to the PEST procedure, which changed the rules for stimulus placement (Pentland, 1980; Watson & Pelli, 1983), came a new category of adaptive procedures – maximum-likelihood procedures (Figure 5). The maximum-likelihood procedure utilizes a large number of hypothetical psychometric functions to determine the probability of obtaining the given responses of the participant. After each trial (i.e. stimulus word in the case of speech testing), the probability of obtaining the responses of the participant to all the trials throughout the duration of the test given each psychometric function, is calculated. The function which yields the highest probability is then used to determine the level of the next stimulus. At the end of the procedure, the most likely psychometric function is used to calculate threshold.

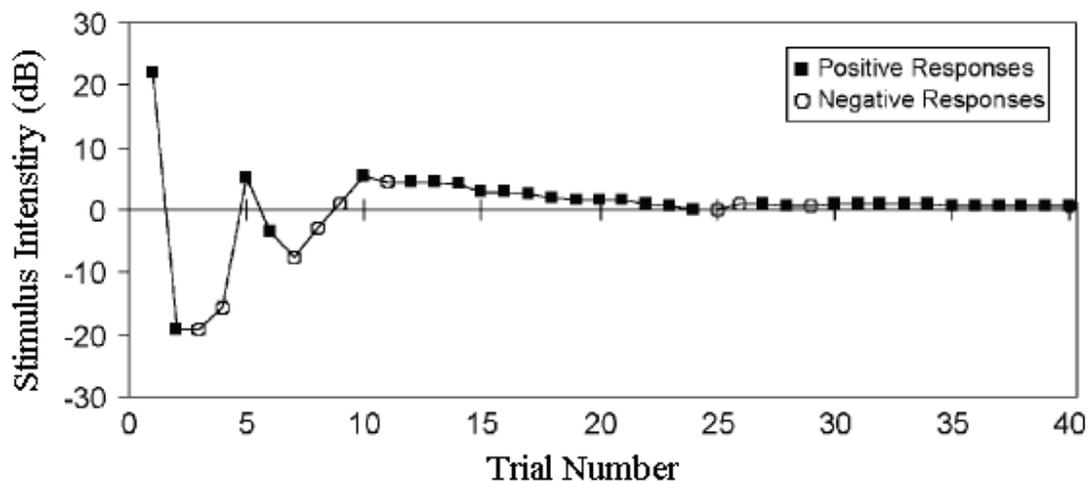


Figure 5. Maximum-likelihood adaptive track (Adapted from Leek, 2001). Decibel values are relative to an arbitrary threshold of 0 dB, which is represented by the horizontal line.

An example of a maximum-likelihood procedure is the QUEST procedure, developed by Watson and Pelli (1983). This procedure utilizes all the information in the preceding trials, as well as prior knowledge (from the literature or previous experiments) to set the next stimulus level, and calculates the final threshold from all the available data in the track.

Research has supported the reliability of thresholds measured using maximum-likelihood methods. In a study using a 2AFC maximum-likelihood procedure for the determination of taste and smell thresholds, test-retest reliability was found to be very good (Linschoten et al.,

2001). Thresholds were also shown to be sensitive to individual differences and stable within individuals over time, with reliability being determined using four robust methods (repeated measures ANOVA, Pearson correlation, differences between successive threshold measurements, and the distribution of standard deviations).

1.5 Comparing Procedures

Comparisons between adaptive procedures and traditional methods of constant stimuli, and between different types of adaptive procedures themselves, can be made with respect to their administration time, accuracy, efficiency and reliability.

Administration time has often been overlooked in previous studies, as it is a direct function of the number of trials used in a particular procedure, and therefore, has an impact on the efficiency of the procedure (which is a more commonly reported measure). Administration time is important in this study, however, as it provides useful information in terms of the procedure's clinical applicability. No matter how efficient a procedure, it must still be able to be completed within an acceptable length of time to be suitable for clinical use, as clinicians have limited time available to obtain results.

The accuracy of a threshold estimate is a useful measure when comparing different psychophysical procedures, as it indicates the suitability of the procedure for the intended purpose and highlights any biases that may be present. The accuracy, however, can be difficult to determine, as in behavioural experiments and tests, the true value of an individual's threshold is often unknown (Treutwein, 1995). In the present study, the degree of consistency between the adaptive threshold estimate and the threshold measured using an agreed-upon technique (the current clinical method of speech audiometry) and a known correlate (the PTA) was used to infer accuracy. As noted by Treutwein (1995), however, the true bias of any procedure can only be evaluated using simulations, and therefore, the behavioural data obtained in this study will only be an indicator of the true degree of accuracy of the adaptive tests.

The term ‘efficiency’, as it is used in this study, is a statistical term, which refers to the precision and tightness of various distributions of equivalent procedures. It can be determined by the number of trials or length of time taken to reach a certain accuracy range of the true threshold value when procedures are different lengths, or by the difference between the obtained threshold estimate and the true threshold when procedures are equal lengths. A single numerical value, known as the ‘sweat factor’ (Taylor & Creelman, 1967) has been used in previous research to compare the efficiency of different procedures that track different percentage correct scores. Because it is determined by the accuracy of the threshold, within-subject variance estimates and the number of trials in a particular procedure, the sweat factor can be used to compare procedures of different lengths that use different rules and parameters.

‘Reliability’ can be used to refer to either test-retest reliability or within-test reliability. Test-retest reliability is a measure of the degree of consistency between an individual’s measured threshold estimates during different testing sessions, while within-test reliability refers to the degree of variance during a testing procedure. In this study, test-retest reliability will be investigated with regard to each speech test, and within-test reliability will contribute to the determination of the optimal termination criterion for use with each adaptive procedure.

1.5.1 Comparison between Adaptive Methods and the Method of Constant Stimuli

Numerous studies have compared adaptive methods with the method of constant stimuli, with the majority finding that adaptive methods have advantages, especially with regard to efficiency (Saber & Green, 1997; Watson & Fitzhugh, 1990).

1.5.1.1 Administration Time of Adaptive Procedures versus Methods of Constant Stimuli

Studies have shown that the administration time of adaptive procedures is considerably less than that of equivalent constant stimuli procedures (Clark & Stewart, 1968; Turpin, McKendrick, Johnson, & Vingrys, 2002). In a test investigating visual field sensitivity, a

maximum-likelihood procedure approximately halved the time required to obtain a threshold when compared to a constant stimuli method, without any loss of accuracy or reliability (Turpin et al., 2002).

With adaptive speech testing, the rate of presentation of the target words is determined by the length of time taken for the participant to make a response, and is therefore, more tailored to the individual participant than conventional speech testing, which uses a recording with preset intervals in which the participant must make a response. Stach, Davis-Thaxton and Jerger (1995) found that a manual mode of speech stimuli presentation, which allows the presentation of speech stimuli to be dictated by the rate of the participant's responses (as is the case with adaptive procedures), reduced administration time by 22%. Obviously this manual mode of stimuli presentation is not specific to adaptive procedures, and could easily be applied to methods of constant stimuli by producing more efficient recordings with very short silent periods which could be controlled by the examiner (e.g. by using the pause button on the C.D. player).

1.5.1.2 Accuracy of Adaptive Procedures versus Methods of Constant Stimuli

The accuracy of thresholds obtained using adaptive methods has generally been shown to be comparable to those obtained using methods of constant stimuli (Buss, Hall, Grose, & Dev, 2001; He, Dubno, & Mills, 1998). In one study investigating simultaneous and backward masking thresholds, both staircase and maximum-likelihood procedures were shown to produce estimates that were not significantly different from those obtained by a method of constant stimuli, which was used as a reference (Buss et al., 2001). Another study showed that a maximum-likelihood procedure used to measure frequency and intensity discrimination gave thresholds that were generally consistent with those obtained from a method of constant stimuli, and were consistent with previous data (He et al., 1998).

Several studies have shown that thresholds estimated using adaptive procedures tend to be slightly lower than those estimated using methods of constant stimuli (Kollmeier, Gilkey, & Sieben, 1988; Saberi & Green, 1997; Taylor, Forbes, & Creelman, 1983). Saberi and Green (1997) showed that interaural timing difference (ITD) thresholds for participants tested using

a method of constant stimuli (both 2AFC and 4AFC) were significantly higher than when tested using both a maximum-likelihood procedure and a two-down, one-up staircase procedure (both 2AFC and 4AFC). Although the difference was statistically significant, however, only three participants took part in the study, which may limit the extent to which the results can be generalized. Taylor, Forbes and Creelman (1983) showed that thresholds obtained from the original PEST adaptive procedure tended to be consistently lower, more stable, and less biased than those obtained from methods of constant stimuli when using a 2AFC task; however, they acknowledge that because the difference was small, it may not be of great practical importance.

In terms of the accuracy of the psychometric function generated from data collected in an adaptive procedure, one may expect that because adaptive procedures tend to place the majority of trials near the target threshold value, the function may not be well-defined at points far removed from this target. Leek, Hanna and Marshall (1992) found that the number of trials was a crucial factor in obtaining an accurate psychometric function. Through the use of computer simulations they showed that tracks consisting of 50 and 100 trials produced slope estimates that were biased high, while tracks of 200 trials or more produced accurate reflections of the true psychometric function.

1.5.1.3 Efficiency of Adaptive Procedures versus Methods of Constant Stimuli

One of the main advantages of adaptive procedures over methods of constant stimuli in the determination of various sensory thresholds is thought to be their increased efficiency (Saber & Green, 1997; Watson & Fitzhugh, 1990), and as such, many researchers refer to this advantage when justifying and explaining the importance of their research regarding adaptive procedures (e.g. Leek et al., 1992). Using simulations, Saber and Green (1997) showed that the method of constant stimuli produced the largest sweat factor, and was, therefore, the least efficient method for determining auditory thresholds when compared to both maximum-likelihood and staircase procedures.

While the vast majority of studies have shown that adaptive methods are more efficient than non-adaptive methods, a few have produced conflicting results, suggesting that methods

of constant stimuli possess equal or greater efficiency when compared to maximum-likelihood procedures (Simpson, 1988) or staircase procedures (McKee, Klein, & Teller, 1985). Simpson (1988) showed through the use of simulations, that the method of constant stimuli was as effective as adaptive methods for the determination of threshold. Watson and Fitzhugh (1990), however, noted that Simpson's simulations were flawed, as they relied on the fact that the difference between a guess and the actual threshold was known, an assumption which is not true when testing real participants.

A further point to note is that there is a reciprocal relationship between accuracy and efficiency (Johnson, Balwantray, & Shapiro, 1992). Increasing the number of trials in an adaptive run will be likely to improve the accuracy of the estimate, but may decrease the efficiency; while decreasing the number of trials will be likely to decrease the accuracy of the estimate, but may increase the efficiency. It is important, therefore, to weigh up the importance of high accuracy and short administration time, in order to determine the procedure with the greatest efficiency for the particular situation in which it is used.

1.5.1.4 Reliability of Adaptive Procedures versus Methods of Constant Stimuli

The reliability of adaptive procedures is often reported to be one of their major advantages over methods of constant stimuli (Leek, 2001; Shelton et al., 1982). Direct comparisons between the reliability of threshold estimates obtained with methods of constant stimuli and adaptive procedures are difficult, however, as differing numbers of trials used in the two procedures can lead to differences in reliability that are not a direct result of the procedure. One study found that speech tests carried out using methods of constant stimuli should contain approximately 450 scorable items for the purposes of optimizing reliability (Gelfand, 1998). If this finding is applied to the conventional speech audiometry test, which only contains 30 scorable items per estimate on the performance-intensity curve, it would seem that the current procedure has less than optimal reliability.

1.5.2 Comparison between Staircase Procedures and Maximum-Likelihood Procedures

As well as comparisons between methods of constant stimuli and adaptive methods, comparisons have also been made between the different categories of adaptive procedures, notably staircase procedures and maximum-likelihood procedures.

1.5.2.1 Administration Time of Staircase Procedures versus Maximum-Likelihood Procedures

The number of trials taken to reach a threshold estimate using staircase and maximum-likelihood procedures is more commonly recorded in research, as opposed to the length of time taken to carry out the procedures. In one study (Formby, Sherlock, & Green, 1996), a significant difference in the length of time taken to complete each procedure was recorded, with an automated, adaptive, maximum-likelihood test of pure-tone threshold taking significantly longer than the conventional staircase technique. The difference was largely due to an increased number of trials being necessary to estimate thresholds in the maximum-likelihood method. In addition, it was noted that a larger amount of time was required to explain the instructions of the maximum-likelihood task to the participants. Prior experience with conventional pure-tone audiometry, however, may have been a confounding variable in this study – as all participants were part of an annual employee hearing conservation programme, it is likely that they had previous experience with conventional pure-tone audiometry. The participants were, therefore, more comfortable with the staircase procedure and did not require such extensive instructions.

1.5.2.2 Accuracy of Staircase Procedures versus Maximum-Likelihood Procedures

Research is divided upon the accuracy of threshold estimates obtained with staircase procedures and maximum-likelihood procedures. While Watson and Fitzhugh (1990) found that a maximum-likelihood procedure produced more stable threshold estimates than a two-down, one-up staircase procedure, a number of studies comparing the two procedures in terms of their determination of various sensory thresholds, have shown that they produce comparable estimates (Buss et al., 2001; Gu & Green, 1994).

In a study carried out by Formby et al. (1996), standard pure-tone audiometry, which employs the use of an adaptive staircase procedure, was compared to an automated, adaptive, maximum-likelihood procedure. The results showed no significant differences between threshold estimates from 500 Hz through to 8 kHz obtained using the two procedures. Similar results have also been found by other researchers using different psychophysical tasks such as intensity discrimination (Marvit, Florentine, & Buus, 2003).

1.5.2.3 Efficiency of Staircase Procedures versus Maximum-Likelihood Procedures

Opinion is divided with regard to the most efficient adaptive procedure. Several studies have shown that there is no significant difference between the efficiency of staircase and maximum-likelihood procedures (Amitay, Irwin, Hawkey, Cowan, & Moore, 2006; Buss et al., 2001), while others have shown that the staircase procedure is more efficient (Marvit et al., 2003) or the maximum-likelihood procedure is more efficient (Shelton & Scarrow, 1984). As different studies used the procedures with different psychophysical tasks and employed different parameters in the procedures, it is difficult to accurately determine if the efficiency of a particular procedure applies to its use in all tasks.

Shelton, Picardi and Green (1982) suggest that different adaptive procedures are useful for measuring certain sensory abilities under different circumstances. The results of their study investigating forward-masking and simultaneous-masking auditory tasks, showed that both staircase and maximum-likelihood procedures produced comparable threshold estimates in a relatively short amount of time (less than 30 trials). They noted, however, that the maximum-likelihood procedure could converge on an accurate threshold after only 10 trials, thus making it the more efficient procedure for gathering threshold information rapidly.

1.5.2.4 Reliability of Staircase Procedures versus Maximum-Likelihood Procedures

In terms of test-retest reliability, studies have shown no significant differences between staircase procedures and maximum-likelihood procedures (Amitay et al., 2006; Formby et al., 1996). In one study which investigated the test-retest reliability of staircase and maximum-

likelihood procedures used with frequency discrimination and backward-masking tasks (Amitay et al., 2006), it was shown that there was a significant learning effect, whereby individuals' scores improved in subsequent testing sessions; however there was no effect of procedure (staircase or maximum-likelihood) or interaction between session and procedure, suggesting that learning did not significantly differ between procedures.

1.6 Optimal Design Criteria and Parameters in Adaptive Procedures

The design criteria and parameters that are applied to adaptive procedures can influence their administration time, accuracy, efficiency and reliability. A selection of the most important design criteria and parameters that apply to the adaptive procedures used in this study – open-set versus closed-set response formats; number of trials; and specific parameters relating to staircase and maximum-likelihood procedures – are discussed below.

1.6.1 Open-Set versus Closed-Set Response Formats

The majority of studies have investigated adaptive procedures coupled to closed-set tasks. Those that have focused on adaptive procedures and open-set tasks have generally shown them to be efficient and reliable in the determination of thresholds (Nilsson, Soli & Sullivan, 1994). Because the number of possible responses is greatly increased in open-set tasks (i.e. they are a function of the individual's capacity for responding), the documented advantages of increased numbers of alternatives in adaptive forced-choice tasks (Amitay et al., 2006; Kollmeier et al., 1988; Schlauch & Rose, 1990; Shelton & Scarrow, 1984) may apply to, and be extended by, the use of an adaptive test with an open-set response format.

Using computer simulations and a behavioural detection-in-noise task to investigate adaptive threshold estimates, Schlauch and Rose (1990) found that as the number of alternatives in an adaptive forced-choice task is increased from two to four, the variability of estimates decreases or remains the same, and the accuracy of the estimate generally increases. Similarly, Amitay et al. (2006) demonstrated the advantages of 3AFC over 2AFC tasks for inexperienced participants.

Several studies have indicated that there is a relationship between the number of alternatives in a forced-choice task and the adaptive rule that is used (Amitay et al., 2006; Kollmeier et al., 1988; Leek et al., 1992; Schlauch & Rose, 1990). Using computer simulations Kollmeier et al. (1988) investigated the efficiency of 2AFC and 3AFC tasks using staircase procedures with both two-down, one-up and three-down, one-up rules. They found that the 2AFC task combined with the two-down, one-up rule was the least efficient and produced the most variability, while the 3AFC task combined with the three-down, one-up rule was the most efficient and produced the least variability. Behavioural data from human participants produced results that tended to support the findings of the computer simulations; however, differences were small and relatively inconsistent, and it was concluded that a researcher should focus more on other experimental design criteria when attempting to increase the efficiency of an adaptive procedure.

1.6.2 Phonemic Scoring versus Whole-Word Scoring

Phonemic scoring, such as that employed by conventional speech audiometry, has a number of advantages over whole-word scoring. It increases the number of scorable items in a test – e.g. the words in the meaningful CVC word lists have three phonemes each, thus phonemic scoring increases the scorable items three-fold compared to whole-word scoring – which increases accuracy and reliability (Gelfand, 1998). In comparison to whole-word scoring, phonemic scoring provides a more direct and precise measure of the reception of the acoustical cues of speech, and results are less affected by non-acoustic factors such as lexical context (Gelfand, 1998).

1.6.3 Starting Level

A study investigating staircase procedures with 4-10 trials for testing visual field sensitivity found that there was no clear relationship between starting level and the accuracy of threshold estimates (Johnson et al., 1992). It was noted that efficiency generally depended on starting level, however, procedures differed with regard to whether it was more efficient to start near threshold or far from threshold (Johnson et al., 1992). Intrinsically, a starting point further from threshold would require more time to reach the vicinity of the threshold, and

therefore, take longer to complete than a procedure that started close to threshold. Potential fatigue of the participant must also be taken into account when determining a starting level, as an increased administration time would undoubtedly lead to greater participant fatigue towards the end of the test, which could have an effect on the accuracy of the obtained threshold estimate.

1.6.4 Target Levels

The majority of early adaptive procedures were developed to track the 50% threshold (e.g. Dixon & Mood, 1948). Green (1990), however, suggests that adaptive procedures that track higher target levels produce more accurate results, as behavioural data from his study showed that the standard deviation of threshold estimates is smaller when tracking the 94% correct level compared to the 70.7% correct level. Ultimately, however, the specific purpose of the procedure will determine what target level is chosen for tracking.

1.6.5 Termination Criteria

Different studies have employed different criteria as a basis to terminate adaptive procedures. The most commonly employed termination criteria include i) stopping after a fixed number of reversals (in staircase procedures) (Bernstein & Gravel, 1990; Wetherill & Levitt, 1965); ii) stopping after a fixed number of trials (King-Smith, 1984; Smallman & MacLeod, 1994; Verdon & Haegerstrom-Portnoy, 1996; Wolfson & Graham, 2000); or iii) stopping once a certain confidence interval has been reached (Treutwein, 1997). While some authors suggest that using a confidence-interval based termination criterion is the most efficient method (Treutwein, 1997), more recent studies have shown that termination criteria based on fixed numbers of trials provide threshold estimates with comparable accuracy to those based on confidence intervals (Alcalá-Quintana & García-Pérez, 2005; Anderson, 2003). Stopping a procedure after a fixed number of trials also has the added advantage of providing certainty with regard to the length of the procedure (Watson & Pelli, 1983). Kontsevich and Tyler (1999) note that this certainty is important in behavioural tests because participants often have difficulty distributing their effort evenly throughout a procedure when the duration of the procedure is variable and unknown.

1.6.6 Number of Trials

Studies have shown that as the number of trials is increased in a staircase procedure, the size of the standard error/variability decreases (Schlauch & Rose, 1990). The sweat factor also decreases up to approximately 100 or 200 trials (Leek et al., 1992). Similarly, studies investigating various sensory abilities in human participants have shown that the variability of results obtained from maximum-likelihood adaptive procedures decreases with increasing trials, but does not decrease significantly when the number of trials is increased beyond a certain point (Florentine et al., 2001; Green, 1993). The number of alternatives in a forced-choice task also impacts upon the number of trials required to obtain a threshold within an acceptable accuracy range. Lam, Dubno & Mills (1999) used computer simulations to show that 40-50 trials are adequate for 3AFC and 4AFC tasks, whereas 2AFC tasks require at least 120 trials to obtain a similarly accurate threshold estimate.

The majority of mathematical models used to test adaptive procedures in the literature tend to predict that the variability of a threshold estimate will decrease and approach zero as the number of trials is increased. Behavioural data using human participants, however, have shown greater variability than that predicted by mathematical models (Kollmeier et al., 1988). While showing a decrease in variability as the number of trials is increased, the value does not approach zero, but rather a non-zero value (Kollmeier et al., 1988). Green (1993) found similar results in an adaptive yes-no pure-tone threshold measurement task, where he noted that variability was larger for human participants than computer simulations.

This increased variability in behavioural data can be explained by a violation of the assumption that an individual's psychometric function remains constant throughout the testing procedure. Hall (1983) suggested that factors such as fatigue, attention, and learning effects may cause an individual's sensitivity to change during the testing procedure, causing the psychometric function to shift. Likewise Green (1993) concluded that the variability in the threshold measurements of the human participants was most likely due to day-to-day fluctuations in their thresholds. A study of minimum gap detection in human participants using a maximum-likelihood procedure supported these findings by showing no decrease in the variability of the minimum gaps detected when the number of trials was increased;

however, when this scenario was replicated in a computer simulation, the variability was reduced with larger numbers of trials (Florentine et al., 2001). In terms of the number of trials necessary to obtain an accurate measure of threshold, it can be concluded that if daily variability in the sensory ability under measurement is larger than the random error caused by the procedure, increasing the number of trials beyond a certain point in an attempt to reduce within-test variance and increase test-retest reliability will not be beneficial.

1.6.7 Parameters Specifically Relating to Staircase Procedures

There is debate regarding the validity of using Levitt's (1971) transformed staircase procedures to target their presumed percentage correct values. García-Pérez (1998) ran large numbers of computer simulations to investigate the effect of procedural characteristics on the convergence of different adaptive staircase procedures with their specified percentage correct values. The result of these simulations did not support the validity of Levitt's (1971) proposed staircase transformations under all procedural conditions; rather, they showed that convergence on the targeted percentage correct value depended more on step size than on the specific staircase rule. The results also indicated that there is an optimal step size ratio (between the step up and the step down) that leads to convergence on the specified target, which differs for different staircase rules, and is unaffected by absolute step size.

With the implementation of the optimal step size ratios in each of the staircase procedures, simulations investigating the effects of trial number, starting value and efficiency were carried out (García-Pérez, 1998). It was found that as long as the relative step size is not less than 0.5, and threshold is estimated from at least 8 reversals, the staircase rules are insensitive to starting value and give unbiased threshold estimates.

The step size variations used in staircase procedures can vary, with some procedures employing constant step sizes throughout (e.g. Swanson, 1996; Voltz & Zanker, 1996), and others making use of a preliminary phase with larger step sizes (e.g. Snowden, Hess, & Waugh, 1995; Turano & Heidenrich, 1996). The rationale for the use of larger step sizes at the beginning of a procedure is that it quickly targets the threshold of interest, whereas the use of constant (smaller) step sizes throughout a procedure leads to a slower approach to the

vicinity of the threshold, but gives participants more time to familiarize themselves with the task. García-Pérez (1998) recommended the abandonment of the preliminary phase with larger step sizes, and instead suggested investing more time performing a longer constant step size procedure. The reasoning behind this idea is that after the preliminary phase with larger step sizes has approached the targeted threshold, the smaller step size tracking procedure provides very little extra information as it generally does not move far from its starting value. Using a constant step size procedure with slightly larger step sizes would, therefore, be more efficient if the time saved by not performing a preliminary phase with larger step sizes is invested running a longer tracking procedure (García-Pérez, 1998).

The number of reversals used to estimate the threshold from an adaptive staircase procedure is typically between 6 and 8 (Wetherill & Levitt, 1965); although there is considerable variation in this number among previous studies (Heidenrich & Turano, 1996; Nicholas, Heywood, & Cowey, 1996). A study by Buss et al. (2001), which involved measuring simultaneous and backward-masking thresholds using Levitt's (1971) three-down, one-up staircase procedure, found that estimates based on two and four reversals were relatively stable when compared to estimates based on six reversals, on average differing by only 0.3 - 0.7 dB in adult participants.

1.6.8 Parameters Specifically Relating to Maximum-Likelihood Procedures

Maximum-likelihood procedures require the use of a baseline psychometric function, with a pre-determined slope (indicating the probability of different thresholds within a population). This initial psychometric function is updated continually for the participant's responses based on Bayes' theorem (Watson & Pelli, 1983). It is not essential, however, that the slope of the initial psychometric function be particularly accurate, as misestimates of its value have been shown to have little effect on the final threshold estimate (Green, 1992).

Maximum-likelihood procedures also require a method for choosing the stimulus intensity for each trial. The original QUEST procedure placed each trial at the current mode of the psychometric function, while later variants of the procedure advocated the use of the mean or median values (King-Smith, Grigsby, Vingrys, Benes, & Supowit, 1994). The initial starting

value chosen for maximum-likelihood procedures has been shown to have little effect on the final threshold estimate (Green, 1992), as the procedure quickly hones in on the vicinity of the targeted threshold according to the responses of the participant.

The final threshold estimate is calculated by applying a maximum-likelihood algorithm to the obtained data from all the trials. This algorithm can be applied independently of the procedure, and therefore, can be used on any set of data whether it was obtained via a maximum-likelihood or staircase procedure.

1.7 Adaptive Speech Tests

Although there has been a great deal of previous research concerning adaptive procedures, there has been limited research carried out with regards to the implementation of adaptive procedures into *speech* tests. Mackie and Dermody (1986) implemented a simple staircase procedure into a forced-choice speech test for children, which they called the Monosyllabic Adaptive Speech Test (MAST). The test was shown to produce reliable and accurate results of children's SRTs, evidenced by the fact that similar results were obtained from a method of constant stimuli, and thresholds correlated to their PTAs.

More recently Zera (2004) implemented an adaptive maximum-likelihood procedure into the modified rhyme test to measure speech intelligibility. In this study, it was found that the adaptive method was an accurate and highly efficient procedure for estimating the speech-to-noise ratio required to obtain different percentage correct scores against a constant level of background noise.

A precursor version of the University of Canterbury Monosyllabic Adaptive Speech Test (UC MAST) program used in this study was previously investigated as a tool to identify children with auditory processing disorders, with the adaptive algorithm being used to adjust the degree of filtering in a filtered words test (McGaffin, 2007). Although the sensitivity and specificity of the test was not able to be assessed, it was found to produce reliable results when administered on repeated testing sessions.

1.8 Aims of this Study

The design of this study included two distinct phases – the Preliminary Testing Phase and the Clinical Testing Phase. The aim of the Preliminary Testing Phase was to compare the administration time, accuracy, efficiency and reliability of different adaptive speech tests (with different adaptive procedures and parameters), and determine the optimal adaptive procedures and parameters for implementation into a closed-set and open-set speech test for use in the clinical Audiology setting. The Clinical Testing Phase of this study aimed to test the optimal adaptive speech tests identified in the Preliminary Testing Phase on a clinical population, and compare the administration time, accuracy and efficiency with that of conventional speech audiometry. Since there is little research in the area of adaptive speech testing, the rationale of the current study was to determine the advantages and disadvantages that an adaptive speech test provides over the method of constant stimuli that is currently employed in clinical speech testing.

1.9 Hypotheses

The following hypotheses were proposed:

- 1) The adaptive speech tests will require significantly less time to complete than conventional speech audiometry (when compared using an equal number of trials), as they involve a manual mode of presentation wherein the rate of presentation of each target word is dictated by the speed of the participant's responses (Stach et al., 1995).
- 2) The thresholds obtained with the adaptive speech tests will show improved or at least comparable accuracy to those obtained with conventional speech audiometry, as the adaptive procedures have increased trial placement around threshold (Buss et al., 2001; Taylor et al., 1983);
- 3) The overall efficiency of the adaptive tests will be greater than that of conventional speech audiometry (Saber & Green, 1997; Watson & Fitzhugh, 1990);
- 4) The thresholds obtained with the adaptive procedures will show greater test-retest reliability than those obtained with conventional speech audiometry (Leek, 2001);

Shelton et al., 1982), as a greater number of trials/scorable items in the vicinity of the targeted threshold will be used to determine the final estimate;

- 5) There will be no significant differences between the optimal staircase and maximum-likelihood adaptive speech tests in terms of their accuracy (Buss et al., 2001; Gu & Green, 1994), efficiency (Amitay et al., 2006; Buss et al., 2001) and reliability (Amitay et al., 2006);
- 6) The adaptive procedures with larger step sizes at the beginning will require significantly less time to reach the vicinity of the targeted threshold than adaptive procedures with constant step sizes throughout (Turano & Heidenrich, 1996); however, there will be no significant difference between the length of time taken to administer the procedures with constant step sizes and larger step sizes at the beginning when a termination criterion dictating a certain level of accuracy is used (García-Pérez, 1998);
- 7) The closed-set adaptive tests will require significantly less time to complete than the open-set adaptive tests, as they do not require an examiner to score responses, but instead are scored instantly by the computer program, thus eliminating the intermediary scoring step;

1.10 Structure of the Study and Thesis Organization

The design of the study is two-fold. The first phase – the Preliminary Testing Phase – involved the implementation of a number of different adaptive procedures and parameters into both closed-set and open-set speech tests. The second phase – the Clinical Testing Phase – involved the testing of the optimal combinations of procedures and parameters identified in the Preliminary Testing Phase on a clinical population.

The overall structure of this thesis is as follows:

- a) Chapter Two contains the rationale for the selection of the procedures and parameters used in the adaptive procedures in the Preliminary Testing Phase of the study;

- b) The general methods that are relevant to both the Preliminary Testing Phase and the Clinical Testing Phase are presented in Chapter Three;
- c) Chapter Four contains the specific methods, results and main findings of the Preliminary Testing Phase of the study, in which the optimal combinations of adaptive procedures and parameters were identified and refined;
- d) Chapter Five contains the specific methods, results and main findings of the Clinical Testing Phase of study, in which the optimal combinations of procedures and parameters identified in the Preliminary Testing Phase were tested on a clinical population;
- e) A general discussion of the findings and conclusions is presented in Chapter Six.

Chapter Two

Selection of Test Materials, Procedures and Parameters for the Preliminary Testing of the Adaptive Tests

2. Selection of Test Materials, Procedures and Parameters for the Preliminary Testing of the Adaptive Tests

The following discussion provides justification for the test materials, procedures and parameters that were chosen for implementation in the adaptive speech tests used in this study.

2.1 Test Materials - Closed-Set and Open-Set Speech Tests

A closed-set speech test comprising the NU-CHIPS word lists (Katz & Elliot, 1978) was used to create the closed-set adaptive speech tests used in this study. Although this speech test was designed for use with children aged 2.5 years and over and is usually implemented as a picture-pointing task, the current study used an adapted version employing written words as opposed to pictures, to make it more suitable for use with adult participants and easier to incorporate into the software. Unlike the majority of previous research, which has focused on adaptive procedures with 2AFC tasks, the present study utilizes a 4AFC task, as the NU-CHIPS test comprises four alternative answers. This greater number of alternatives has advantages in that it reduces the effects of chance responding, decreases variability and increases accuracy of the threshold estimate (Schlauch & Rose, 1990).

As there is more limited data concerning open-set adaptive procedures employing phonemic scoring systems, the section of the study implementing adaptive procedures into open-set speech tests acted as a pilot. Boothroyd and Nitttrouer's (1988) meaningful consonant-vowel-consonant (CVC) words were chosen as the stimuli for the adaptive open-set speech tests. As these were the same stimuli as those used in conventional speech audiometry, a direct comparison between the results obtained from the two methods (adaptive and non-adaptive open-set methods) was possible.

2.2 Selection of Adaptive Procedures for Implementation into the Closed-Set Speech Tests

2.2.1 Staircase Procedure

As the majority of previous studies have utilized adaptive staircase procedures that were designed for and used in conjunction with 2AFC or yes/no tasks, the validity of using unmodified versions of these procedures in conjunction with a 4AFC is questionable. The specific parameters under which a particular staircase rule was formulated dictates whether it is suitable for use in other types of tasks, such as those with a greater number of forced-choice alternatives.

In order to determine an individual's speech reception threshold on a 4AFC task, the use of Dixon and Mood's (1948) simple staircase procedure is not valid. Threshold, as it is defined in this study, is the intensity level at which the probability of a correct response is half way between chance performance and perfect performance (Kaernbach, 2001). Dixon and Mood's (1948) simple staircase rule (one-up, one-down) converges on the 50% correct level, which, in the case of forced-choice tasks, is not halfway between chance and perfect performance. Using the simple staircase rule does not take into account the effects of chance responding, and therefore comparing the level at which 50% is obtained using a forced-choice procedure and the 50% correct score obtained with conventional open-set speech audiometry would be erroneous. In the case of the adaptive closed-set tests utilizing the NU-CHIPS word lists with four alternative answers, chance performance is 25% and perfect performance is 100%. The threshold level, therefore, is 62.5% - halfway between 25% and 100% - and thus, an adaptive procedure targeting the 62.5% level is required in order to make a valid comparison between the estimated threshold and the SRT obtained with conventional speech audiometry.

Levitt's (1971) transformed staircase rules were designed for 2AFC tasks and are unable to target the 62.5% level. Additionally, the optimal step size ratios for use with transformed staircase rules set out by García-Pérez (1998) are only suggested for use with 2AFC tasks, as the stability of trial placement around a targeted threshold breaks down when used with

3AFC and 4AFC tasks (García-Pérez & Alcalà-Quintana, 2005). Kaernbach's (1991) weighted rules, that are able to target any performance level on a psychometric function, are, therefore, most suitable for implementation in the 4AFC closed-set tests used in this study. As well as being able to target the desired threshold level, weighted staircase rules have also been shown to be slightly better overall - in terms of threshold accuracy and variability - than comparable transformed staircase rules (García-Pérez & Alcalà-Quintana, 2005).

Using the weighted staircase rule, Kaernbach (1991) defines the equilibrium condition for convergence on a point (X_p) as:

$$S_{\text{down}} p = S_{\text{up}}(1-p) \quad [1]$$

Using the above equation¹, the size of the upwards step (S_{up}) and the size of the downwards step (S_{down}) for convergence on the 62.5% correct level ($X_{62.5}$) can be calculated. The rule for this situation would read: Decrease the intensity 0.6 steps after each correct response and increase the intensity 1 step after each incorrect response. This particular version of Kaernbach's (1991) weighted staircase rule was chosen for implementation in the adaptive closed-set speech tests used in this study.

2.2.2 Maximum-Likelihood Procedure

To better define the psychometric function, it is beneficial to obtain information about more than simply the threshold. Targeting a point higher than threshold on the psychometric curve is advantageous, as research has indicated that the reliability of higher points is better (Green, 1990), and when combined with the threshold estimate, can give information about the slope of the psychometric curve, which cannot be derived from a single threshold estimate alone.

The original QUEST procedure (Watson & Pelli, 1983) was chosen as the maximum-likelihood procedure for use in this study primarily because of the availability of documented

¹ This equation uses modified terms that differ from those used by Kaernbach (1991).

parameters for its implementation. In addition to the fact that the procedure is well-documented in the literature, it has also been shown to be efficient in its acquisition of information (Watson & Fitzhugh, 1990), and therefore, provides a potential method for use in the determination of speech recognition information in a clinical Audiology setting.

As well as performing an independent QUEST procedure, it is also possible to use the QUEST algorithm to calculate a higher point on the psychometric function based on the information obtained from an adaptive staircase procedure. That is, the staircase procedure can be used to determine the stimulus level of each trial, while the QUEST algorithm can be used to calculate a final threshold estimate based on the staircase information. As such, the QUEST algorithm was programmed to calculate a higher threshold (82% threshold) during each staircase procedure, in order to obtain added information about the slope of the performance-intensity speech curve.

2.3 Selection of Adaptive Procedures for Implementation into the Open-Set Speech Tests

In the case of an open-set test, threshold is 50%, as this is halfway between chance performance (considered 0%) and perfect performance (100%). Dixon and Mood's (1948) simple staircase procedure, targeting the 50% threshold, is therefore suitable for use in this context. In order to make a valid comparison between the thresholds obtained with the adaptive open-set procedures and conventional speech audiometry, however, the same phonemic scoring system had to be used with each method. A variation of Kaernbach's (1991) weighted rule was therefore applied to the simple staircase procedure, to allow the use of phonemic scoring as opposed to whole-word scoring. This variation allowed the number of phonemes correct in the participants' responses to determine both the direction and size of the following step (see Table 1 on page 38 for the absolute intensity changes corresponding to the number of phonemes correct).

2.4 Selection of Other Parameters for use in the Adaptive Staircase Procedures

2.4.1 Starting Level

The presentation level of the first trial word in all adaptive staircase procedures was chosen to be 40 dB above the participant's PTA. This ensured that the word was clearly audible and gave participants an opportunity to familiarize themselves with the nature of the task before the threshold seeking procedure began (García-Pérez, 1998; Green, Richards, & Forrest, 1989).

2.4.2 Step Size – Constant versus Larger at the Beginning

For both the closed-set weighted staircase procedures and the open-set simple staircase procedures, two different step size variations were employed. The first version of each procedure employed constant step sizes throughout the entire test. In this context, 'constant' refers to increments and decrements in level of a constant size throughout the test (although these increments and decrements are not necessarily equal, as is the case in the closed-set weighted staircase procedures). The second version of each procedure employed larger initial step sizes at the beginning of each test. This larger step size is referred to as the 'initial step size', while the step size for the remainder of the test is referred to as the 'working step size' (see Figure 7 on page 41). This study's implementations of staircase procedures with larger step sizes at the beginning followed that of McGaffin (2007), with changes between the initial and working step size programmed to occur after four reversals.

The rationale for using the two different step size procedures in this study was that each variation provided specific advantages, and a comparison between the two would determine which provided the greatest benefits for use in a clinical speech test. Using larger step sizes at the beginning of an adaptive run is advantageous as the procedure is able to quickly reach the vicinity of the target threshold level, so that when the working increment is employed to fine tune the estimate, it is already operating in the approximate region of the threshold. Using constant step sizes throughout an adaptive test provides more data points at the upper end of

the psychometric function so that a more accurate curve can be plotted, and allows the participant to gain more practice at making responses to trials above threshold.

2.4.3 Initial and Working Step Sizes – Absolute Values

For both the closed-set and open-set adaptive procedures, the working step size was set to a maximum of 2 dB, while the initial step size (only utilized in the procedures with larger step sizes at the beginning) was set to a maximum of 5 dB. In the closed-set adaptive tests, which utilized a weighted staircase rule, the initial step up size corresponded to 5 dB and the initial step down size corresponded to 3 dB (0.6 of 5 dB), while the working step up size corresponded to 2 dB and the working step down size corresponded to 1.2 dB (0.6 of 2 dB). These steps are discussed further in Section 3.3. In the open-set adaptive tests, which utilized a modified simple staircase procedure, step sizes depended on the number of phonemes the participant was able to repeat correctly. With respect to the initial step size, each phoneme corresponded to an increment or decrement of 1.67 dB; while with respect to the working step size, each phoneme corresponded to 0.67 dB. The absolute values for the initial and working step sizes and the corresponding number of phonemes correct that led to the initiation of each step are displayed in Table 1 below.

Table 1. Absolute values for the initial and working step sizes of the adaptive open-set speech tests based on the number of phonemes correct

Phonemes correct	Initial step size (dB)	Working step size (dB)
3 of 3	-5	-2
2 of 3	-1.67	-0.67
1 of 3	+1.67	+0.67
0 of 3	+5	+2

2.4.4 Termination Criteria

Each adaptive staircase procedure employed a 22 reversal termination criterion, whereby the test terminated after 22 reversals about the participant's estimated threshold. Using such a large number of reversals as a termination criterion was chosen as it allowed retrospective analyses of the results, whereby procedures could be terminated at different points using

different termination criteria, and the threshold estimates obtained at each of these points could be analyzed and compared. Termination criteria with a 5 dB accuracy span (± 2.5 dB, $+3/-2$ dB and $+4/-1$ dB of the final threshold estimate) were chosen for comparison (see Figure 6), as 5 dB is the smallest intensity increment used in clinical Audiology practice, and therefore represents adequate accuracy for use in a clinical test. Two of the three termination criteria ($+3/-2$ dB and $+4/-1$ dB of the final threshold estimate) included a larger margin of error on the upper side of the threshold estimate, as the early tests in the Preliminary Testing Phase showed that threshold estimates tended to decrease with increasing numbers of trials. This observation was due to the fact that a supra-threshold starting level was employed, which meant that more trials were presented above, rather than below, each participant's threshold.

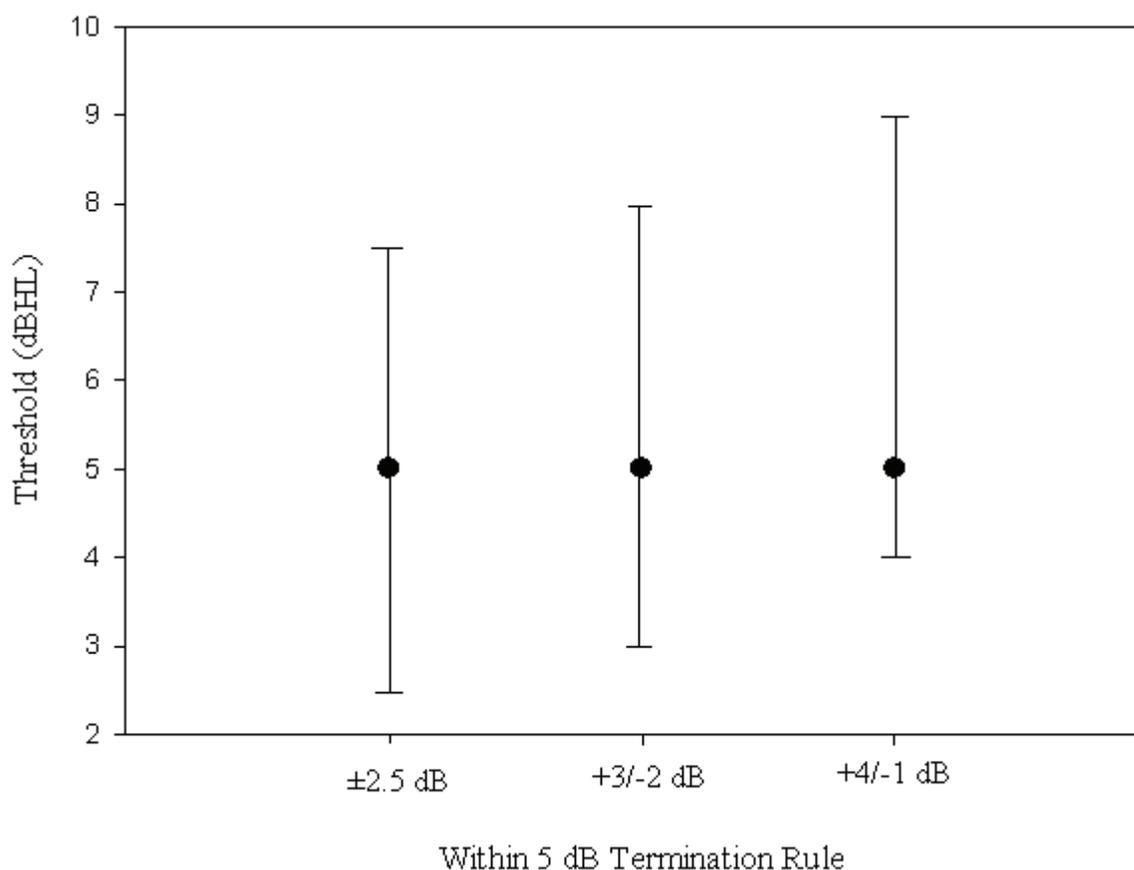


Figure 6. Visual representation of the three 'within 5 dB of the final estimate' termination criteria, displaying the range of accuracy for the threshold of an individual with a final threshold estimate of 5 dB HL.

The program used to run the adaptive tests also calculated the 99% confidence interval for the final threshold estimate of each adaptive procedure. This feature allowed the retrospective analysis of a 99% confidence interval termination criterion, whereby the number of trials and time required to reach this level of accuracy could be compared between tests.

2.4.5 Practice Reversals and Real Reversals – Threshold Calculation

The reversals in the adaptive staircase procedures consisted of both ‘practice reversals’ and ‘real reversals’ (see Figure 7). The practice reversals were included to allow participants a chance to familiarize themselves with the task, without the initial results having an impact on the final threshold estimate. Practice reversals consisted of the first four reversals in all the adaptive staircase procedures and comprised the initial step size trials in the procedures with larger step sizes at the beginning. Threshold was calculated by averaging the midpoints of the final 18 reversals (i.e. the real reversals). The exclusion of the practice reversals from the threshold calculations served two purposes. Firstly, in both constant and larger step size procedures the early practice reversals may not have been as accurate as those in later trials, and therefore their exclusion potentially increased the accuracy of the final threshold estimate. Secondly, in the procedures with larger step sizes at the beginning, it was important not to include the reversals obtained from the initial step sizes in the determination of threshold, as in order to obtain a valid threshold estimate, it must be computed over the range in which the step size remained constant (García-Pérez, 1998).

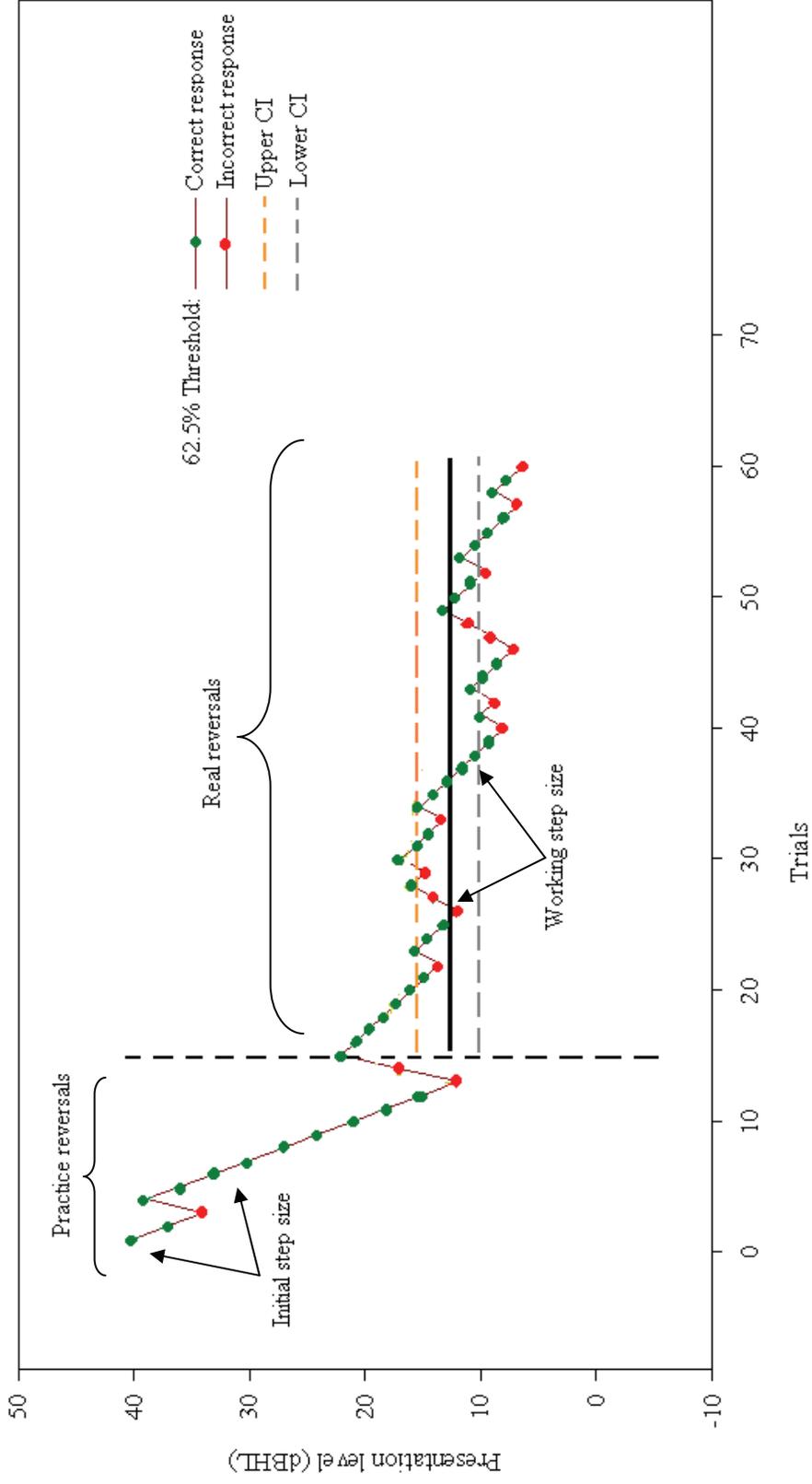


Figure 7. An example of an adaptive run utilized in the closed-set adaptive speech test, showing initial and working step sizes and practice and real reversals. As shown, the final threshold estimate is calculated from taking the average of the midpoints of the final 18 reversals that utilize the working step size. The 99% upper and lower confidence intervals (CIs) are shown as dashed lines.

2.5 Selection of General Parameters for use in the Adaptive Maximum-Likelihood Procedure – QUEST

The QUEST algorithm used in this study was based on the version distributed as part of the Psychophysics Toolbox MATLAB extensions (Brainard, 1997; Pelli, 1997). The 2004 version of this MATLAB extension was ported to LabVIEW by Christina Starfinger and Greg O’Beirne and was incorporated into the UC MAST software.

The QUEST algorithm assumes a Weibull psychometric function (Weibull, 1951), which may be written as a function of dB intensity as follows:

$$W_T(x) = 1 - (1 - \gamma) \cdot \exp\left[-10^{(\beta/20)(x-T+\epsilon)}\right] \quad [2]$$

$$p_T(x) = \min\left[(1 - \delta), W_T(x)\right] \quad [3]$$

T is the point on the psychometric function to be used as threshold, x is the stimulus intensity in dB, and ϵ is an offset introduced so that T will be the testing point where the ‘ideal sweat factor’ is least. β is a slope parameter and γ represents chance performance or the probability of success at zero intensity (which is 0.25 or 25% for a 4AFC task). Equation 3 accounts for lapses on behalf of the participant by reducing the maximum score possible by a factor δ (such that $\delta = 0.01$ limits the maximum score to 99%).

The QUEST routine used in this study was initialized to assume a Weibull psychometric function with the following default settings: $\beta = 3.50$; $\gamma = 0.25$; $\delta = 0.01$; and $p_{\text{Threshold}} = 0.82$.

To better match the threshold range used for the UC MAST program to that of the QUEST algorithm, both the guessed threshold and guessed standard deviation were divided by 10 before being passed to the QUEST routine. The output was then re-scaled by the same factor. This resulted in an effective β of 0.35 being used for the Weibull function algorithm. Ideally, the slope of the Weibull function used in the QUEST algorithm should match the slope of the actual psychometric function being tested (i.e. the performance-intensity curve for speech recognition). In practice, however, Pelli (2005; 2006) suggests that most values of β will

work, but that using the wrong β will normally only decrease the efficiency with which the threshold is estimated without the introduction of significant bias. As shown in Figure 8, the slope of the psychometric function for a typical open-set performance-intensity speech curve is approximately $\beta = 0.84$, but how this compared to a closed-set test (where γ is non-zero) was not known until after the completion of this study. As will be discussed later, pooled data from the adaptive closed-set constant step size procedures of six normal participants indicates that a higher β of 4.75 is most suitable.

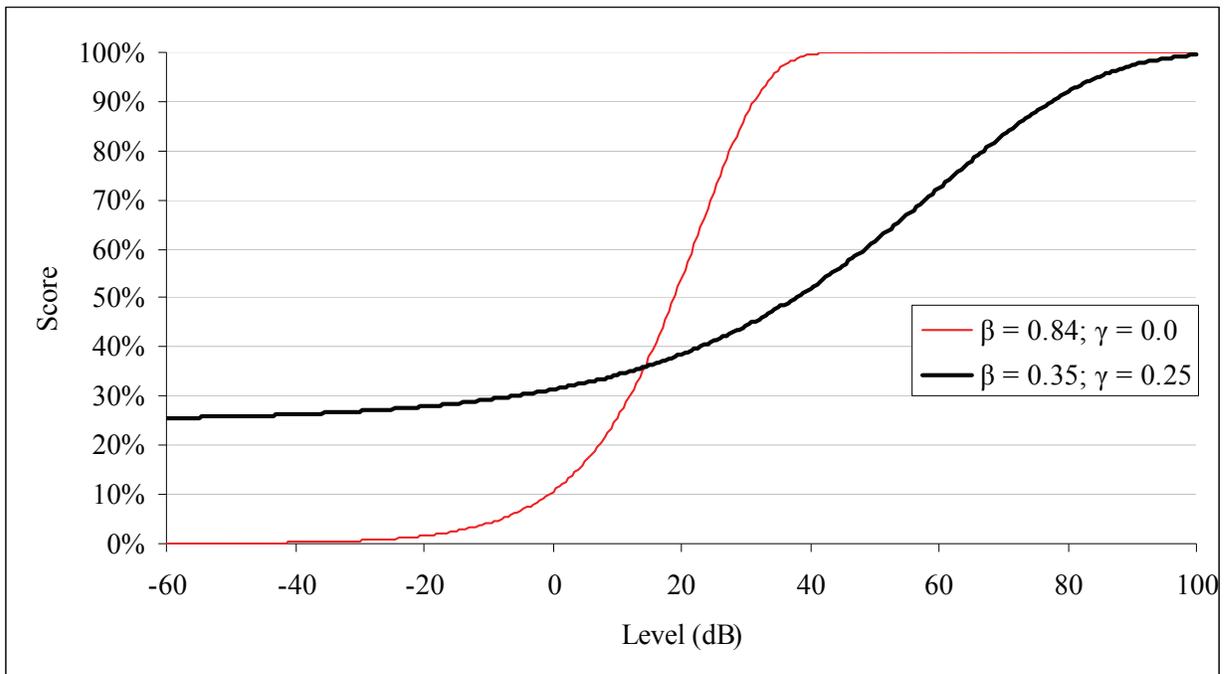


Figure 8. Comparison of the slope of the psychometric function used in the QUEST algorithm (black) to a psychometric function fitted to a typical performance-intensity curve for normal speech audiometry (red). [Speech test data used for red curve: 100% at 45 dB, 87% at 30 dB, and 38% at 15 dB].

The choice of 0.82 as the pThreshold for 2AFC tasks is based on the fact that $p_T(x) = 0.816$ when $(T + \epsilon) = x$ and $\gamma = 0.5$ (Shadlen, 2002). In this case, substituting $\gamma = 0.25$ for a 4AFC produces a $p_T(x) = 0.724$, indicating that the sweat factor is least when pThreshold = 0.724. Since the other parameters for the QUEST routine were documented in combination with a pThreshold of 0.816, it was decided that it would be more appropriate to use the 82% threshold as opposed to the 72% threshold and sacrifice the potential for increased efficiency.

Additionally, the fact that the QUEST procedure was always utilized in a closed-set task meant that the resultant 82% threshold estimate could not be directly compared with the 82% correct threshold obtained using conventional open-set speech audiometry. In order to make a valid comparison, therefore, the QUEST 82% threshold had to be compared to its equivalent percentage correct threshold of conventional open-set speech audiometry, which was calculated to be 76% using equation 4 below.

$$\frac{\text{pThreshold (\%)} - \text{chance performance (\%)}}{\text{perfect performance (\%)} - \text{chance performance (\%)}} \quad [4]$$

$$\frac{82 - 25}{100 - 25} = \frac{57}{75} = 76\%$$

As in the staircase procedures, the starting level of the QUEST procedures was set to 40 dB above each participant's PTA to ensure the initial stimulus word was clearly audible. The termination criterion was set to 50 trials, as this was deemed a suitable length for a clinical test.

Chapter Three

General Methods

3. General Methods

The general methods described in this section relate to both the preliminary and clinical phases of the study, except where otherwise stated. Additional methodological issues and procedures that were employed in the preliminary and clinical phases can be found in Chapters Four and Five respectively.

3.1 Participants

Participation in this study was voluntary, with participants being recruited subject to availability for testing. Prior to inclusion in the study, each participant was administered a short questionnaire (see Appendix I), which sought the following self-reported information:

- Participant's age;
- Participant's visual acuity – gauged by whether the individual could comfortably see a test word on a screen at a distance of approximately 50 cm;
- Participant's degree of literacy – gauged by whether the individual suffered from any known literacy problems;
- Participant's expressive language skills – gauged by whether the individual suffered from any known speech problems;
- Participant's proficiency in the English language – based on self-report.

The following criteria, which could be ascertained from the participants' responses on the questionnaire, were used to exclude any individuals who were not suitable to take part in the study:

- Poor eyesight, to the degree that the individual would be unable to clearly see the visually presented words for the closed-set tests;
- Poor literacy skills, to the degree that the individual would not be able to read the visually presented words for the closed-set tests;
- Poor expressive English language skills, to the degree that the examiner would not be able to accurately score the individual's responses to the open-set tests.

3.2 Equipment

All testing took place in a double-walled, sound-treated room at the University of Canterbury Speech and Hearing Clinic. A Toshiba laptop with a 16-bit soundcard was used to run the adaptive speech program and present the speech stimuli. A twin RCA to 3.5 mm cable connected the sound card of the laptop to the left and right channel inputs of a GSI-61 audiometer (Grason-stadler Corp., USA). An external Elo ET1715L 17" touch screen monitor (Tyco Electronics Corp., USA), connected to the laptop was used to visually present stimulus words to the participants in the closed-set adaptive speech tests. Schematic representations of the test set-up for the adaptive closed-set tests and adaptive open-set tests are displayed in Figures 9 and 10 respectively.

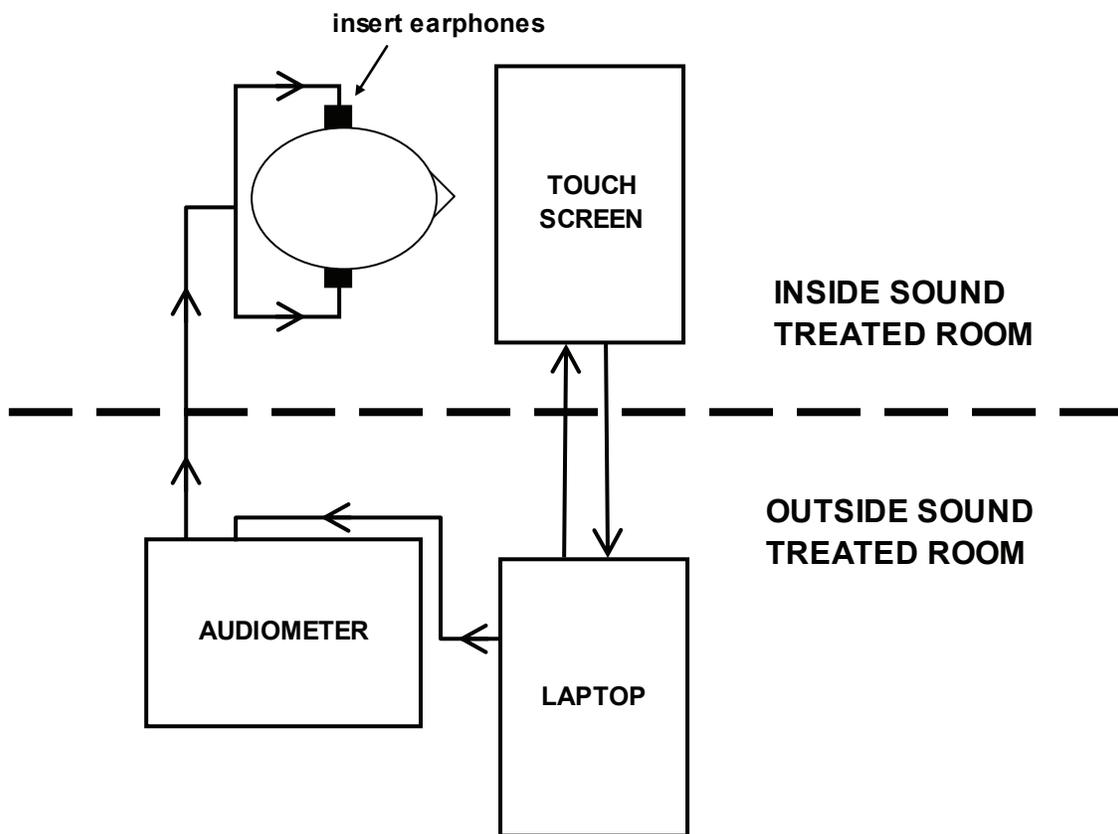


Figure 9. Schematic representation of the set-up of equipment for the adaptive closed-set speech tests.

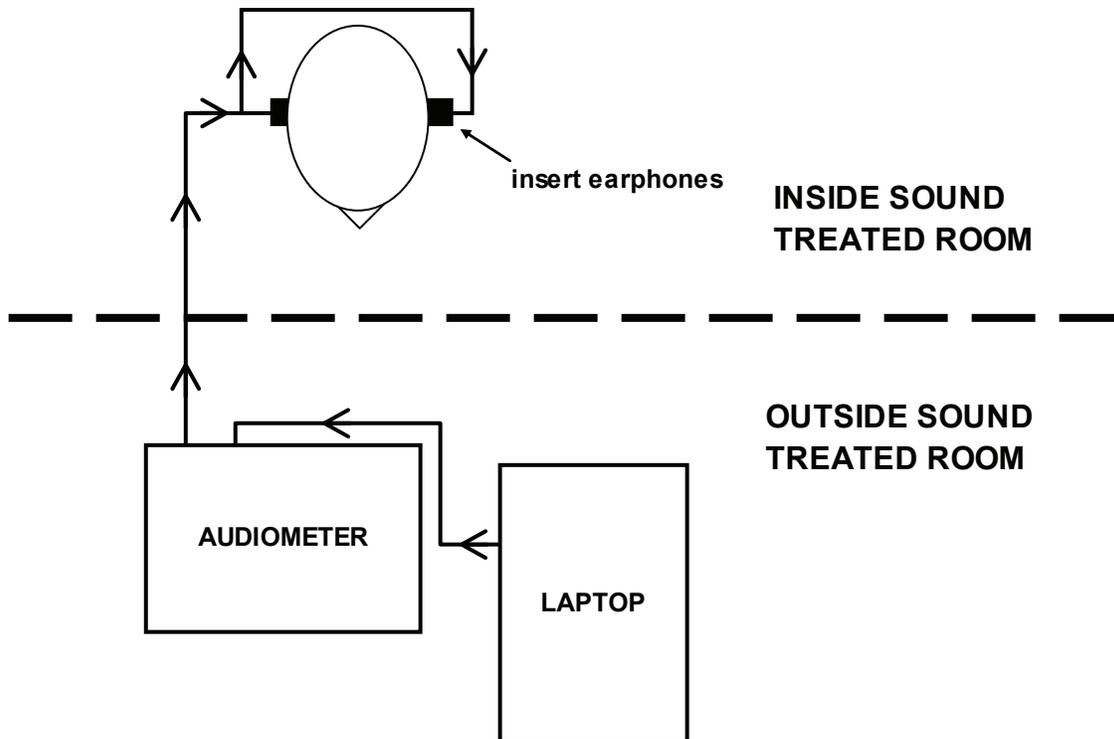


Figure 10. Schematic representation of the set-up of equipment for the adaptive open-set speech tests.

3.2.1 UC MAST Software

The adaptive speech test program – University of Canterbury Monosyllabic Adaptive Speech Test (UC MAST) – was developed by Dr Greg O’Beirne using National Instruments LabVIEW 8.20. The program controlled the presentation of the stimuli in the adaptive speech tests and recorded data – e.g. trial numbers, presentation levels, responses, confidence intervals, provisional threshold estimates and final threshold estimates - to tab delimited text files. An example of the tracking results produced by each adaptive test administered by the UC MAST program is displayed in Appendix III. Prior to test administration, the examiner specified the response mode (closed-set versus open-set), the staircase rule (simple, weighted, transformed), step size variations (constant versus larger at the beginning) and absolute step size (see Figure 11). All other parameters were left on their default settings. The graphical user interfaces of the UC MAST program displayed the alternative responses for the adaptive closed-set tests (panels A and B of Figure 12) and the scoring screen for the adaptive open-set tests (panels C and D of Figure 12).

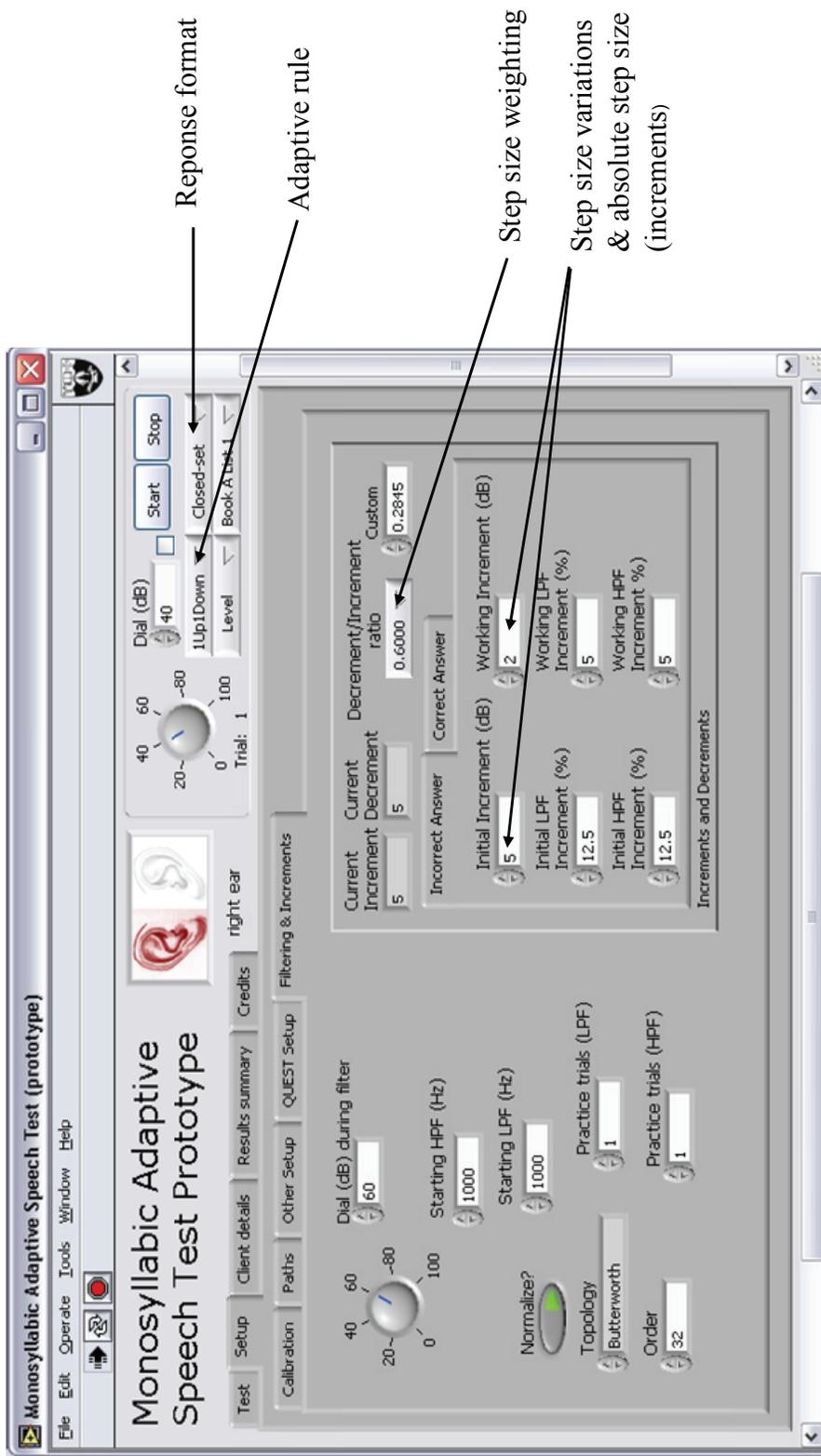


Figure 11. Set-up interface for the UC MAST program.

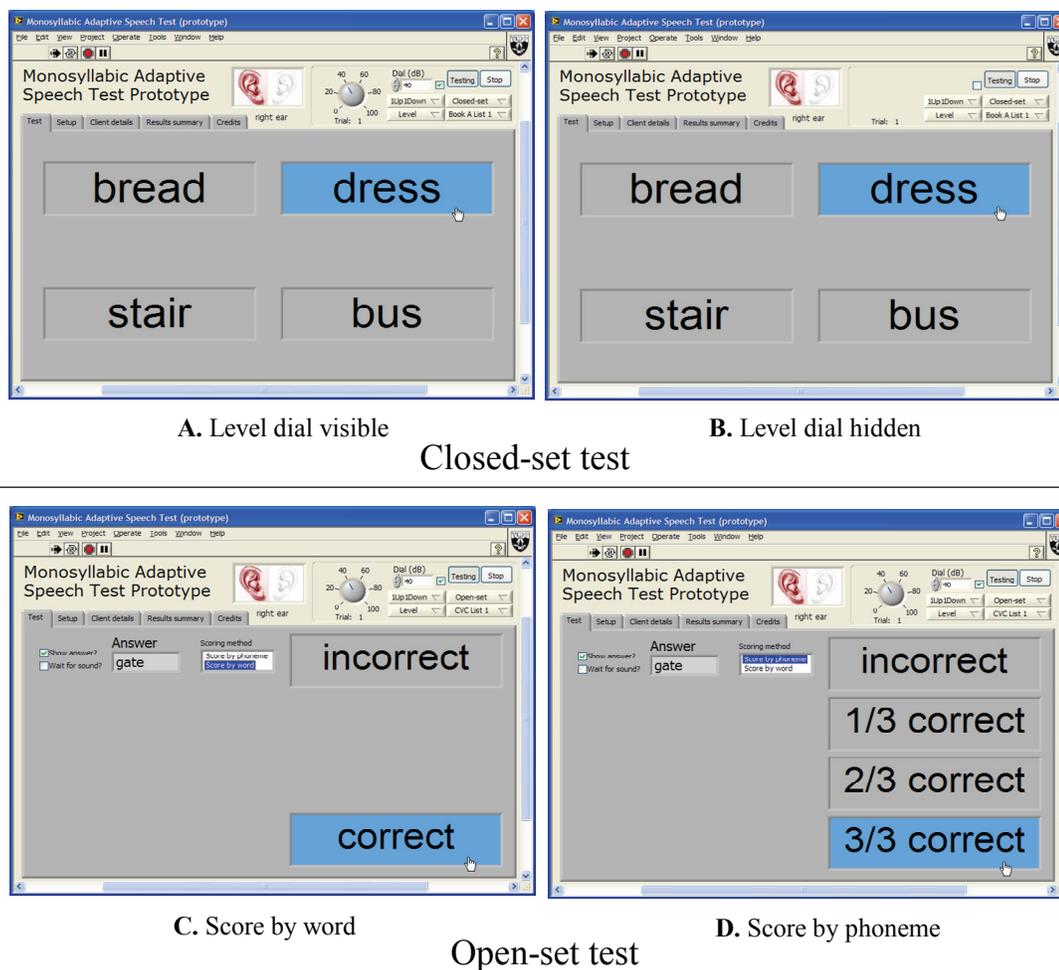


Figure 12. Example of the graphical user interfaces for the UC MAST program under different operating modes.

3.2.2 Calibration of Equipment

Prior to administration of the speech tests, the output signal level of the laptop and the compact disc player were calibrated using a 1 kHz tone to ensure that stimuli from each device were presented at the same signal levels. The output of the calibration tone on the compact disc was adjusted via an attenuator to match the output of the laptop. To ensure an adequate range of presentation levels were available for testing, the UC MAST program was calibrated at 70 dB HL for all participants with a PTA of less than or equal to 30 dB HL (as shown in Figure 13), and 90 dB HL for those with a PTA of greater than 30 dB HL.

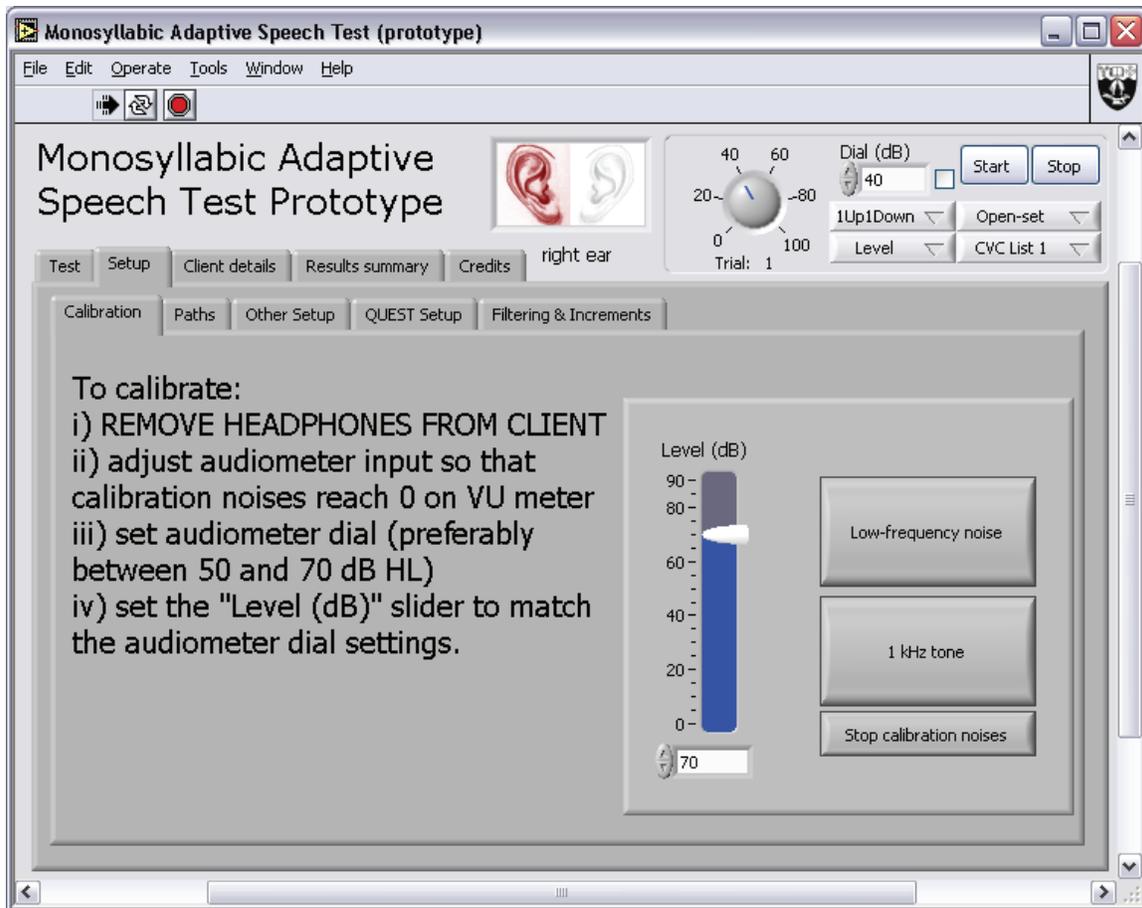


Figure 13. Calibration interface for the UC MAST program

3.3 Test Procedure

Otoscopy was performed to ensure insert earphones were appropriate for use. Pure-tone audiometry at octave intervals from 250 Hz – 8000 Hz was administered using the modified Hughson-Westlake procedure. In cases where thresholds from consecutive octave frequencies differed by 20 dB HL or more, inter-octave frequency testing was also performed.

Four basic categories of speech tests were carried out – non-adaptive open-set and closed-set tests, and adaptive open-set and closed-set tests.

I. Non-Adaptive Open-Set Test (Conventional Speech Audiometry): Three lists of ten monosyllabic words were presented monaurally via insert earphones. The presentation

level of the first list of words, which aimed to give a result close to 100%, was 30 – 40 dB above the average of the participant’s pure-tone thresholds at 1 kHz and 4 kHz. The second list of words was presented at a level 15 – 20 dB below the presentation level of the first list, and the final list was presented at a level 10 – 15 dB below that of the second list. The exact levels chosen were left to the discretion of the examiner, who tested all participants. Each test word was preceded by the carrier phrase “Say”, and followed by a short gap in which participants were required to repeat back the word that they heard. As shown in Table 2, a phonemic scoring system was used to score the participants’ responses, with 0 points awarded for no correct phonemes, 3 points for one correct phoneme, 7 points for two correct phonemes and 10 points for repeating the entire word correctly. The results of each list of ten words at a single presentation level yielded a percentage correct score, which was used to plot a performance-intensity curve on a standard Audiological recording sheet.

II. Adaptive Open-Set Tests (*Staircase procedures with constant step sizes and larger step sizes at the beginning*): The meaningful CVC word lists were presented monaurally via insert earphones. Each test word was preceded by the carrier phrase “Say”. The participant was required to repeat back each word. A manual phonemic scoring system was used to score responses, whereby the examiner selected the icon with the corresponding number of phonemes correct (0 correct, 3 out of 10 correct, 7 out of 10 correct, all correct) on the scoring screen on the laptop (shown in panel D of Figure 12 on page 51). This phonemic scoring system was similar to that used in conventional speech audiometry, but resulted in the adaptive step being multiplied by a factor between +1 and -1, as shown in Table 2, which depended on the number of correct phonemes in the previous target word. Upon the examiner allocating a score to the participant’s response, the next stimulus word was acoustically presented at a level determined by the adaptive algorithm which was based on the participant’s score on the previous word.

Table 2. Comparison of the scoring protocols for conventional speech audiometry and the adaptive open-set tests implemented in this study.

Phonemes correct	Conventional speech audiometry marks awarded	Open-set adaptive test step multiplier
3 of 3	10	1
2 of 3	7	+ $\frac{1}{3}$
1 of 3	3	- $\frac{1}{3}$
0 of 3	0	-1

III. Adaptive Closed-Set Tests (Staircase procedures with constant step sizes and larger step sizes at the beginning, and QUEST): Test words from the NU-CHIPS word lists were presented monaurally via insert earphones. No carrier phrase was presented. Participants were required to choose the word from four alternatives (shown in panels A and B of Figure 12 on page 51) on a touch screen that corresponded to the word they heard. The program used a whole-word scoring system to automatically score responses. Upon making a response, the next target word was presented at a level determined by the adaptive algorithm, and the procedure continued.

IV. Non-Adaptive Closed-Set Test (Conventional NU-CHIPS): Two lists of twenty test words from the NU-CHIPS word lists were presented monaurally via insert earphones – one list at the 82% level determined by the independent QUEST procedure and one list at the SRT level determined by the adaptive closed-set test with constant step sizes. No carrier phrase was presented. The participant used a booklet with four alternative words on each page to point to the word that corresponded to the target word during the pause that followed each test word. A manual whole-word scoring system was used by the examiner to score responses. The final score was converted to a percentage correct score.

Speech tests were administered in random order, based on a random number generator, to eliminate the effects of test order on results. A standard set of instructions (Appendix II) was read to participants for each category of speech test prior to administration of the first speech test in that category.

3.4 Measures

The administration time of instructions and procedures, the accuracy of threshold estimates, and the efficiency and reliability of the tests were compared to determine the optimal closed-set and open-set adaptive speech tests in the Preliminary Testing Phase, and to investigate the advantages and disadvantages of these ‘optimal’ procedures when compared to conventional speech audiometry in the Clinical Testing Phase.

3.4.1 Administration Time

Each speech test was subdivided into two key components – administration of instructions and testing procedure – and the time taken to complete each component was measured separately. The length of time taken to administer instructions for each category of speech test and complete the testing procedures for the non-adaptive speech tests was measured using a stopwatch, while the time taken for participants to carry out the computer-based testing procedures (comprising all the adaptive tests) was independently measured by the UC MAST program itself, which recorded the time at each response made.

I. Administration of Instructions (Preliminary Phase only)

The time taken for the examiner to read the standard set of instructions for each category of speech test (Appendix II) to each participant was recorded. Because the instructions for both versions of the adaptive open-set test (constant step sizes and larger step sizes at the beginning) and all the adaptive closed-set tests (constant step sizes, larger step sizes at the beginning and QUEST) were the same, these were only read once at the beginning of the first test to which they applied, and a brief reminder was given, if required, when subsequent tests were performed. The length of time taken to complete this component of the procedure also included any time spent clarifying instructions or answering questions regarding the task.

II. Testing Procedure

The time taken to complete each speech test procedure was recorded. This time began once the first stimulus was presented and terminated after the participant had

made a response to the final stimulus. The UC MAST program recorded the time at each individual response and the corresponding confidence intervals and provisional threshold estimates, enabling retrospective analyses of the results using different termination criteria. The time taken to complete each speech test using 22 reversals, the ‘within 5 dB of the final threshold estimate’ termination criteria (± 2.5 dB, $+3/-2$ dB and $+4/-1$ dB), and the ‘within the 99% confidence interval of the final threshold estimate’ termination criterion were analyzed.

3.4.2 Accuracy

The accuracy of the threshold estimates obtained with the adaptive speech tests was inferred from their degree of consistency with thresholds obtained from the non-adaptive speech tests. The SRT based on conventional speech audiometry was calculated using a Solver routine based on a Weibull function as the performance-intensity curve. The degree of consistency between the threshold results of each speech test and the participants’ PTA was also taken into account when determining the accuracy of the SRT.

3.4.3 Efficiency

The efficiency of the adaptive staircase procedures was determined by comparing the number of trials and time taken for each participant’s threshold estimate to reach the same degree of accuracy across all the tests. An external indication of accuracy was required as the final threshold estimate obtained from each adaptive staircase procedure would differ between the different tests. As the PTA is a single value and a known correlate of the SRT it was chosen as an indication of test accuracy. The number of trials and the time taken for the threshold estimate of each participant on each adaptive staircase procedure to reach the same decibel level from their PTA was compared. To ensure comparisons could be made between all staircase tests, the decibel level furthest from the PTA (either within ± 6 dB of the PTA – indicative of good agreement – or $\pm 7 - \pm 12$ dB of the PTA – indicative of fair agreement) was used as a constant measure of accuracy. Using this method, however, meant that the efficiency of the QUEST procedure and conventional speech audiometry could not be calculated.

3.4.4 Reliability (Preliminary Testing Phase only)

30% of participants were re-tested on all speech tests (except the non-adaptive closed-set procedure) at least two weeks after their initial testing session. The degree of consistency between the scores obtained with the same tests on both testing sessions was used to give an indication of the test-retest reliability of each procedure.

3.5 Data Analysis

All statistical analyses were carried out using SigmaStat Version 3.11. One-way repeated measures analyses of variance (RM ANOVAs) were performed to determine if there was a significant difference in the time taken to administer the instructions for the different categories of speech tests, and the number of trials and time taken to complete the testing procedures for conventional speech audiometry and the adaptive speech tests (using 22 reversals and the 99% confidence interval as termination criteria). Two-way RM ANOVAs were used to determine if there was a significant difference between the number of trials and the administration time required for completion of the speech tests when using the three different 'within 5 dB of the final estimate' termination criteria for the adaptive speech tests. Pearson product moment correlations were run for the SRT estimate and the 76% and 82% thresholds obtained with each applicable test, with one-way RM ANOVAs also being performed to determine if there were significant differences between the threshold estimates obtained from any of the tests. For those participants in the Preliminary Testing Phase who were re-tested, thresholds obtained on the first and second testing sessions were compared using a Pearson product moment correlation, to determine the test-retest reliability of each test.

Chapter Four
**Preliminary Testing Phase for the Refinement
of the Adaptive Procedures**

4. Preliminary Testing Phase for the Refinement of the Adaptive Procedures

4.1 Introduction

The Preliminary Testing Phase of this study was carried out to determine the optimal procedures and parameters for implementation into a closed-set speech test and an open-set speech test for use in a clinical setting. The findings from this preliminary phase were used to determine which adaptive speech tests would be administered in the Clinical Testing Phase of the study.

4.2 Methods

The General Methods described in Chapter Three relate directly to this phase of the study. In addition, the characteristics of the participants and the specific speech tests that were administered to each participant are set out in the following sections.

4.2.1 Participants

Twenty adult participants (18 years and over) with normal hearing bilaterally were recruited from the University of Canterbury to take part in the preliminary phase of this study. The participants consisted of 7 males and 13 females, and ranged in age from 18-54 years, with an average age of 29.2 years (*S.D.* = 12.2). ‘Normal hearing’ was defined as hearing thresholds of less than or equal to 20 dB HL at octave intervals from 250 Hz to 8 kHz bilaterally. Hearing sensitivity, represented by the pure-tone average (PTA) of hearing thresholds at 500, 1000 and 2000 Hz ranged from -5.0 to 16.7 dB, with an average PTA of 5.3 dB (*S.D.* = 5.3).

4.2.2 Speech Tests

Seven speech tests were administered monaurally to the right ear of each participant. These speech tests comprised:

1. Non-Adaptive Open-Set Test (conventional speech audiometry);
2. Adaptive Open-Set Test (simple staircase rule) - constant step sizes throughout the test;

3. Adaptive Open-Set Test (simple staircase rule) - larger step sizes at the beginning of the test;
4. Adaptive Closed-Set Test (weighted staircase rule) – constant step sizes throughout the test;
5. Adaptive Closed-Set Test (weighted staircase rule) – larger step sizes at the beginning of the test;
6. Adaptive Closed-Set Test (maximum-likelihood rule) – QUEST;
7. Non-Adaptive Closed-Set Test (standard NU-CHIPS).

The order in which speech tests 1 - 6 were administered was randomized to reduce any effects of test order on results. Due to the fact that the presentation levels corresponding to the SRT (determined using the adaptive closed-set test with constant step sizes) and the 82% threshold (determined using the independent QUEST procedure) had to be estimated prior to administration of the standard NU-CHIPS test, this non-adaptive closed-set test was only able to be administered after the other adaptive closed-set tests had been performed. For convenience, this was always the last test to be performed with each participant.

4.3 Results

The administration time, accuracy, efficiency and reliability of results for the adaptive staircase procedures and adaptive maximum-likelihood procedures are, for the most part, presented separately in this section, in order to focus on the determination of the optimal procedures and parameters for each category of adaptive test. For more detailed comparisons between staircase and maximum-likelihood procedures refer to Chapter Five.

4.3.1 Administration Time

The administration time of the instructions and testing procedures were measured separately, as discussed in the General Methods section.

4.3.1.1 Administration of Instructions

Results of a one-way RM ANOVA on Ranks revealed a significant test category effect on the length of time taken to administer instructions (Chi-square = 46.90, $df = 3$, $p < 0.001$). As shown in Figure 14, results of post-hoc pairwise multiple comparison procedures using the Tukey test revealed that the two open-set tests were associated with significantly longer instruction administration time than the two closed-set tests ($p < 0.05$) and the comparison between adaptive and non-adaptive procedures was not significantly different for both open-set and closed-set procedures. Note: Error bars indicate standard deviations.

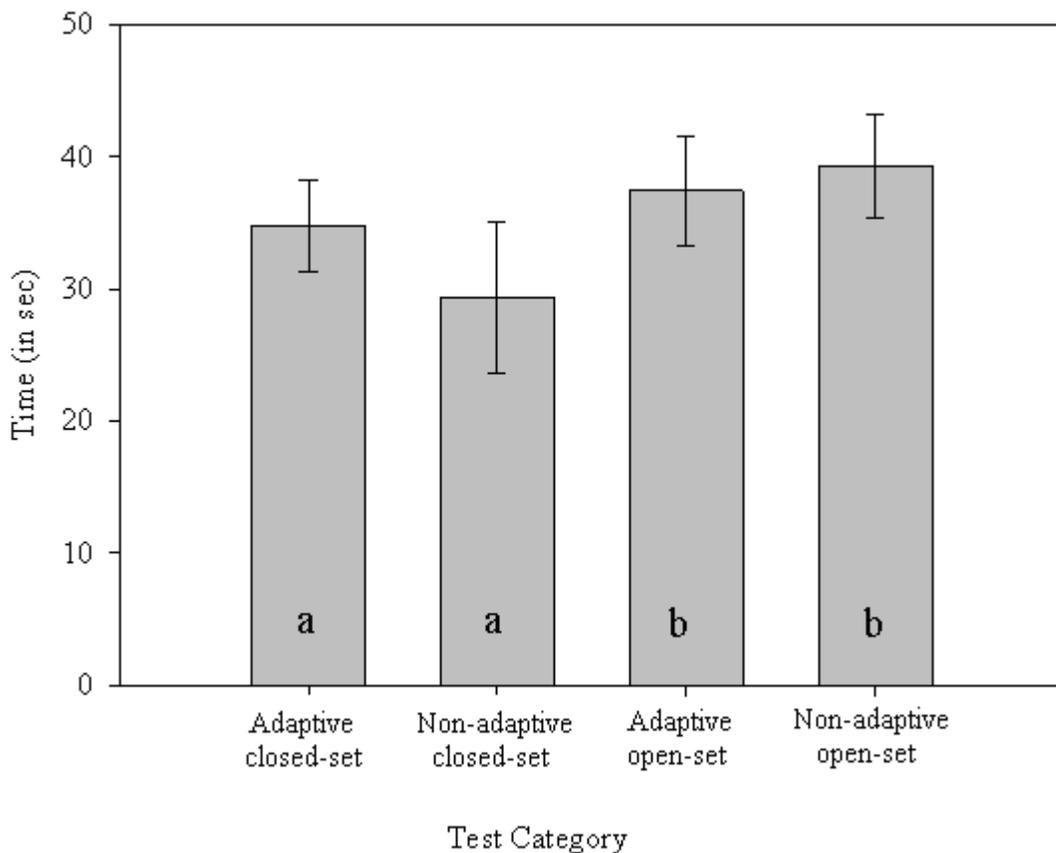


Figure 14. Mean length of time taken to administer instructions for the different categories of speech tests. Group means that are significantly different from each other are labelled with different letters.

4.3.1.2 Administration of Testing Procedure – Staircase Speech Tests

22 Reversal Termination Criterion:

Results of a one-way RM ANOVA on Ranks revealed a significant test effect on the number of trials required to complete each speech test when using 22 reversals as a termination criterion for the adaptive tests (Chi-square = 58.005, $df = 4$, $p < 0.001$). In comparison with conventional speech audiometry, all the adaptive staircase tests (except the closed-set test with larger step sizes at the beginning) required a significantly higher number of trials for test completion (see Figure 15).

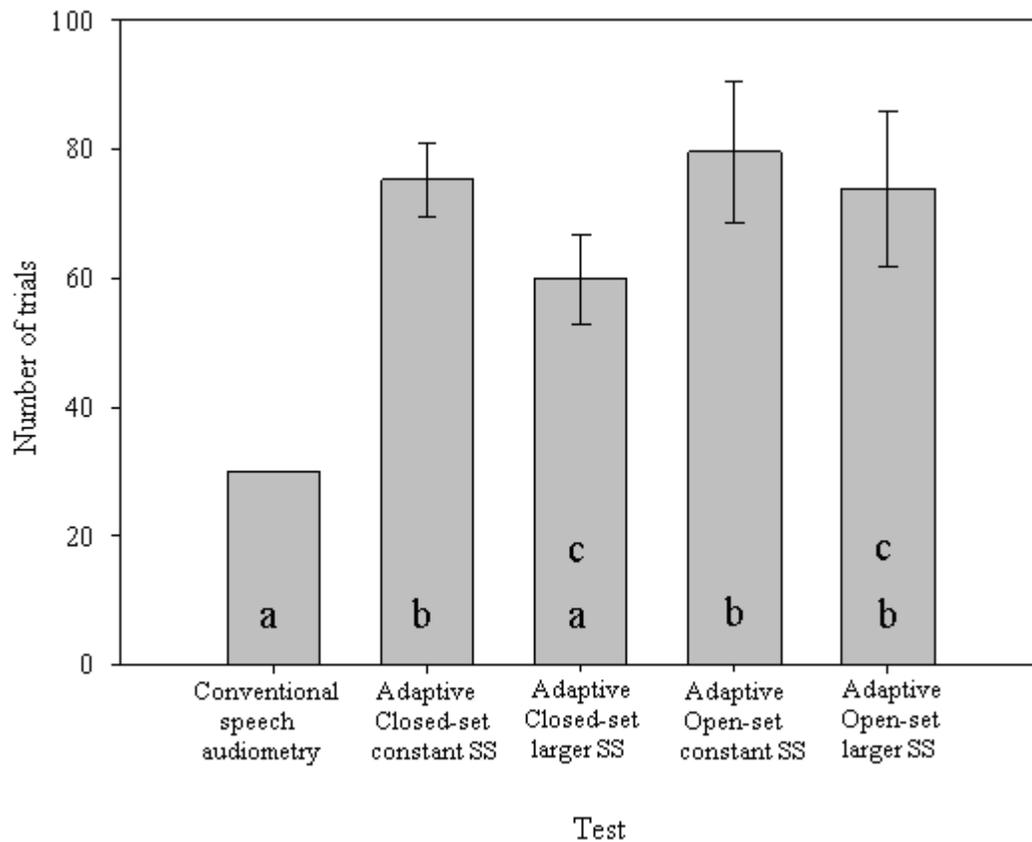


Figure 15. Mean number of trials taken to complete conventional speech audiometry and the adaptive staircase speech tests when using 22 reversals as a termination criterion for the adaptive tests (SS = step size). Group means that are significantly different from each other are labelled with different letters.

Results of a one-way RM ANOVA on Ranks revealed a significant test effect on the time required to complete each speech test when using 22 reversals as a termination criterion for the adaptive tests (Chi-square = 65.560, $df = 4$, $p < 0.001$). As shown in Figure 16, both open-set adaptive staircase tests took a significantly longer length of time to complete than both closed-set adaptive staircase tests and conventional speech audiometry ($p < 0.05$) while no significant time difference was found between conventional speech audiometry and the two closed-set adaptive staircase tests.

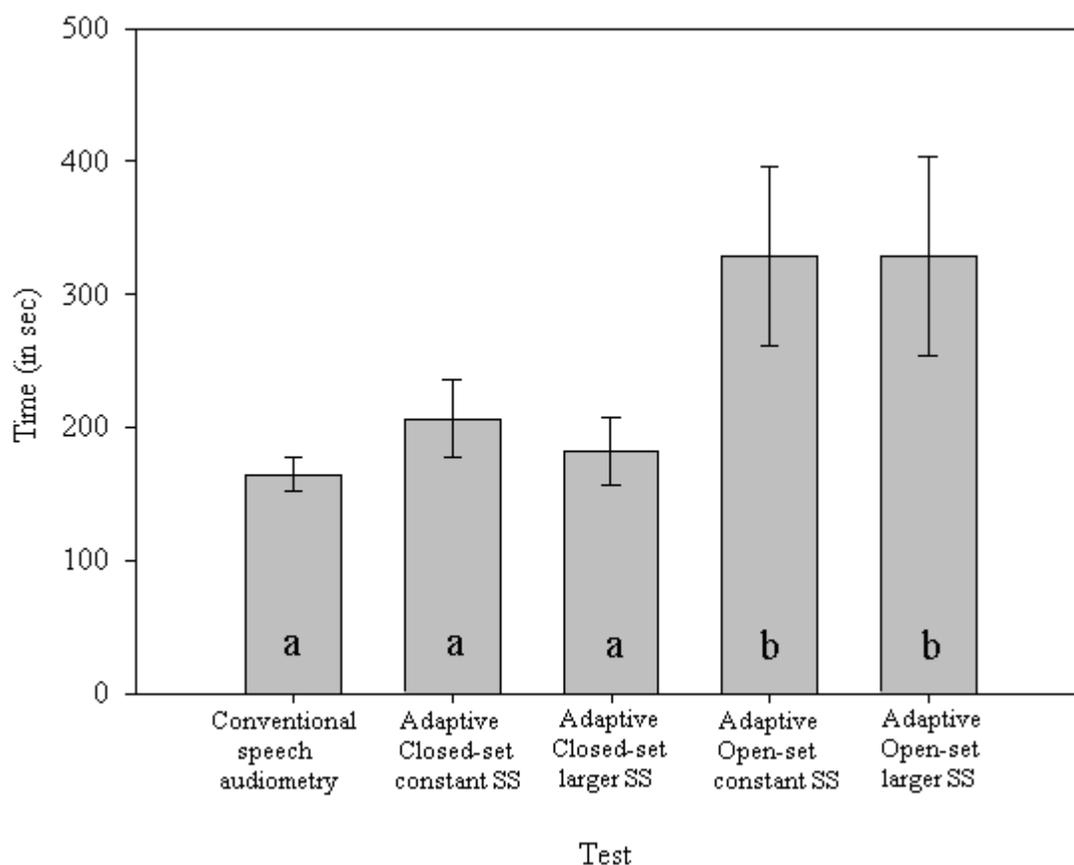


Figure 16. Mean length of time taken to complete conventional speech audiometry and the adaptive staircase speech tests when using 22 reversals as a termination criterion for the adaptive tests. Group means that are significantly different from each other are labelled with different letters.

A two-way RM ANOVA on Ranks revealed a significant test effect [$F(3,114) = 16.195$, $p < 0.001$], and test by termination criterion interaction effect [$F(6,114) = 4.067$, $p < 0.001$], but no significant termination criterion effect [$F(2,114) = 1.989$, $p = 0.151$] on the number of trials required to complete each speech test when using the ‘within 5 dB of the final threshold estimate’ termination criteria for the adaptive tests. As shown in Figure 17, there was no significant difference between the number of trials required to complete the adaptive open-set tests, the adaptive closed-set test with larger step sizes at the beginning and conventional speech audiometry when using any of the three ‘within 5 dB of the final threshold estimate’ termination criteria. The adaptive closed-set test with constant step sizes required a significantly greater number of trials to complete than conventional speech audiometry and the adaptive closed-set test with larger step sizes at the beginning when using the equivalent termination criterion.

'Within 5 dB of the Final Threshold Estimate' Termination Criteria:

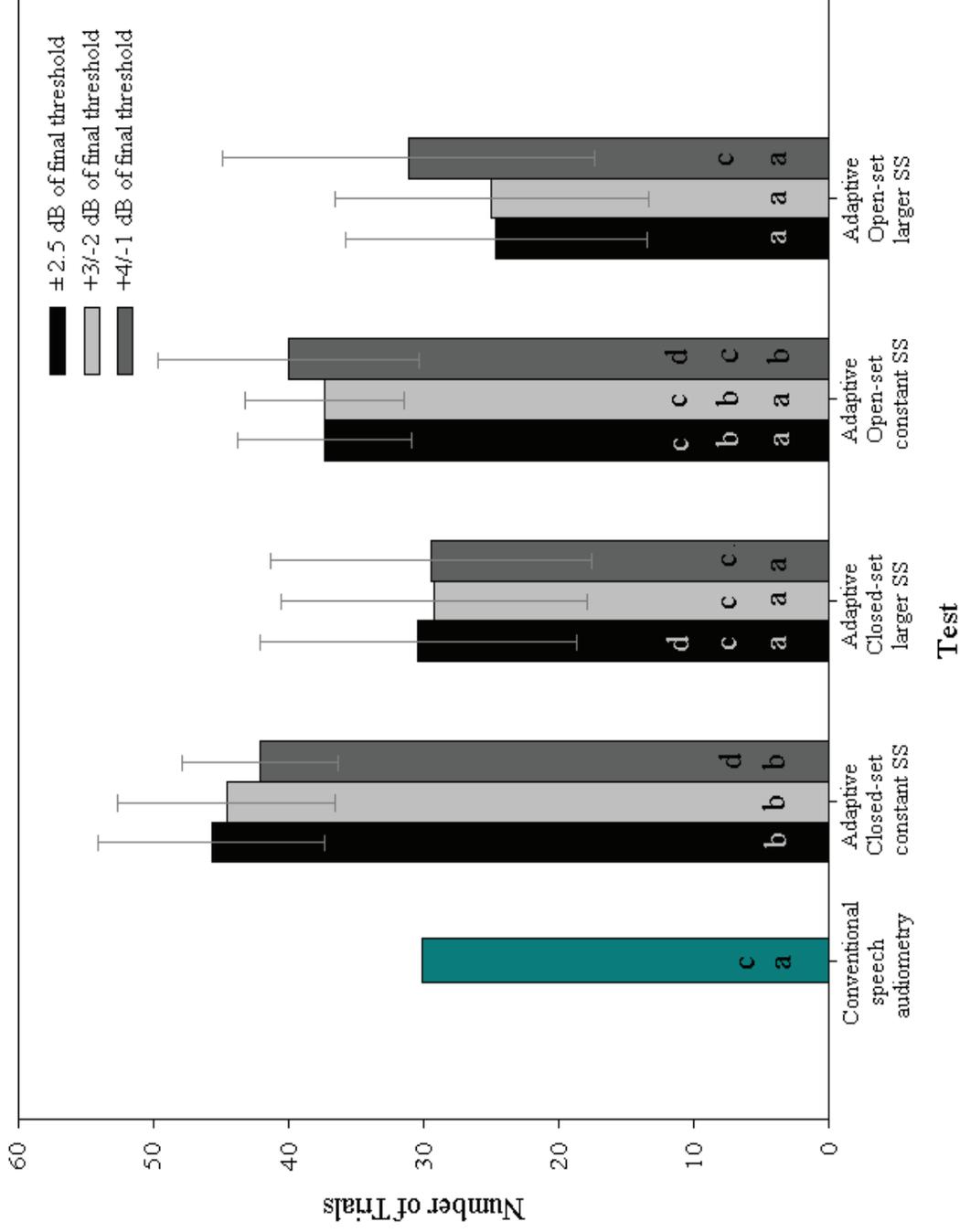


Figure 17. Mean number of trials required to reach within 5 dB of the final threshold estimate in each adaptive staircase speech test, using the termination criteria ± 2.5 dB, $+3/-2$ dB and $+4/-1$ dB of the final threshold estimate. The mean number of trials required to complete conventional speech audiometry is included as a reference. Group means that are significantly different from each other are labelled with different letters.

A two-way RM ANOVA on Ranks revealed a significant test effect [$F(3,114) = 9.566$, $p < 0.001$], a significant termination criterion effect [$F(2,114) = 4.122$, $p = 0.024$], and a significant test by termination criterion interaction effect [$F(6,114) = 4.327$, $p < 0.001$] on the time required to complete each speech test when using the ‘within 5 dB of the final threshold estimate’ termination criteria for the adaptive tests. As shown in Figure 18, there was no significant time difference between the two closed-set adaptive tests; however, both required significantly less time to complete than conventional speech audiometry. The open-set adaptive test with larger step sizes at the beginning took a significantly shorter amount of time to complete than conventional speech audiometry and the equivalent constant step size test when using the ± 2.5 dB and $+3/-2$ dB of the final threshold estimate termination criteria.

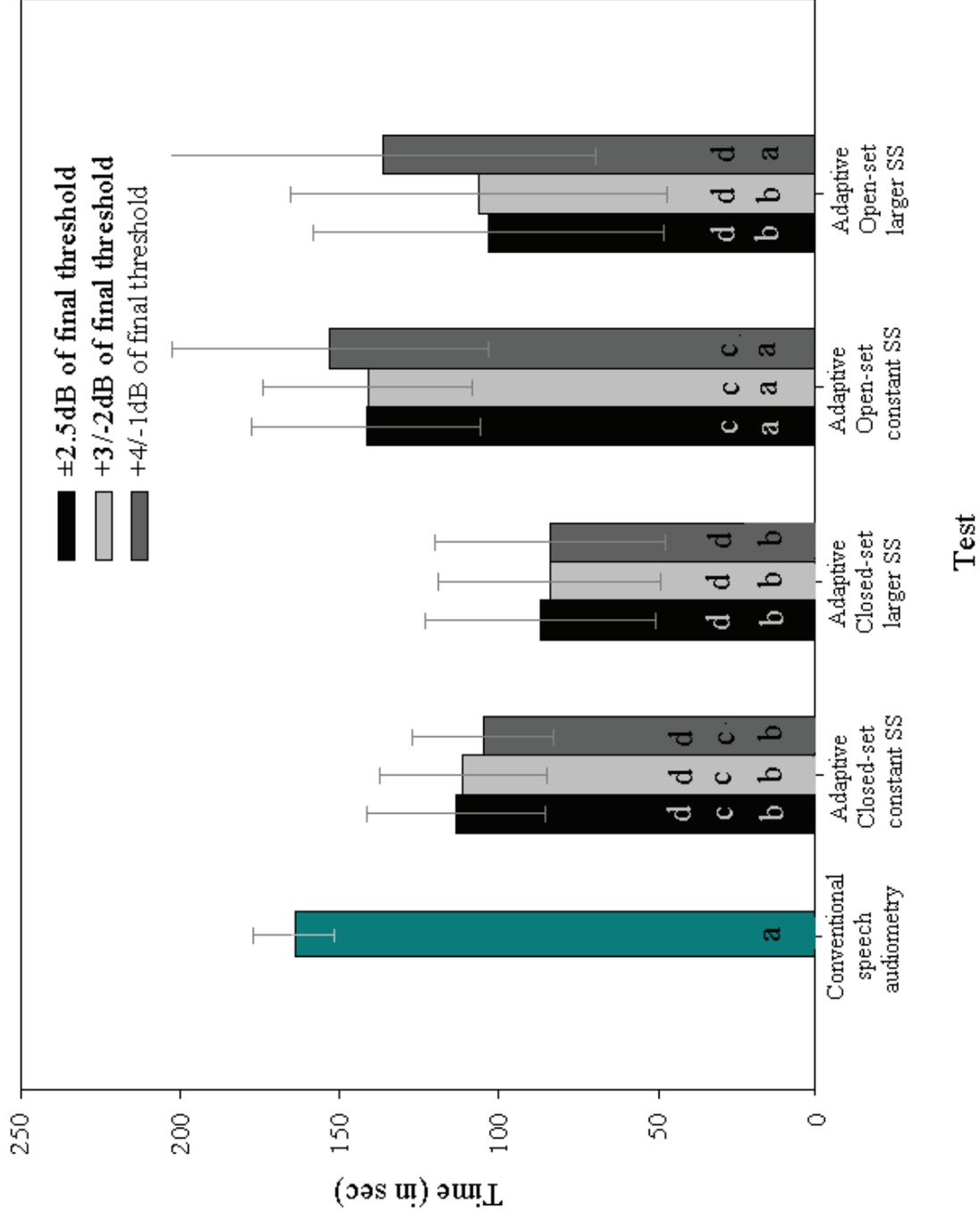


Figure 18. Mean length of time required to reach within 5 dB of the final threshold estimate in each adaptive staircase speech test, using the termination criteria ±2.5 dB, +3/-2 dB and +4/-1 dB of the final threshold estimate. The mean length of time taken to complete conventional speech audiometry is included as a reference. Group means that are significantly different from each other are labelled with different letters.

99% Confidence Interval Termination Criterion:

Results of a one-way RM ANOVA on Ranks revealed a significant test effect on the number of trials required to reach a threshold within the 99% confidence interval of the final threshold estimate on the adaptive staircase tests (Chi-square = 35.566, $df = 4$, $p < 0.001$). All the adaptive tests, except the closed-set test with larger step sizes at the beginning, required a significantly larger number of trials to complete than conventional speech audiometry (Figure 19).

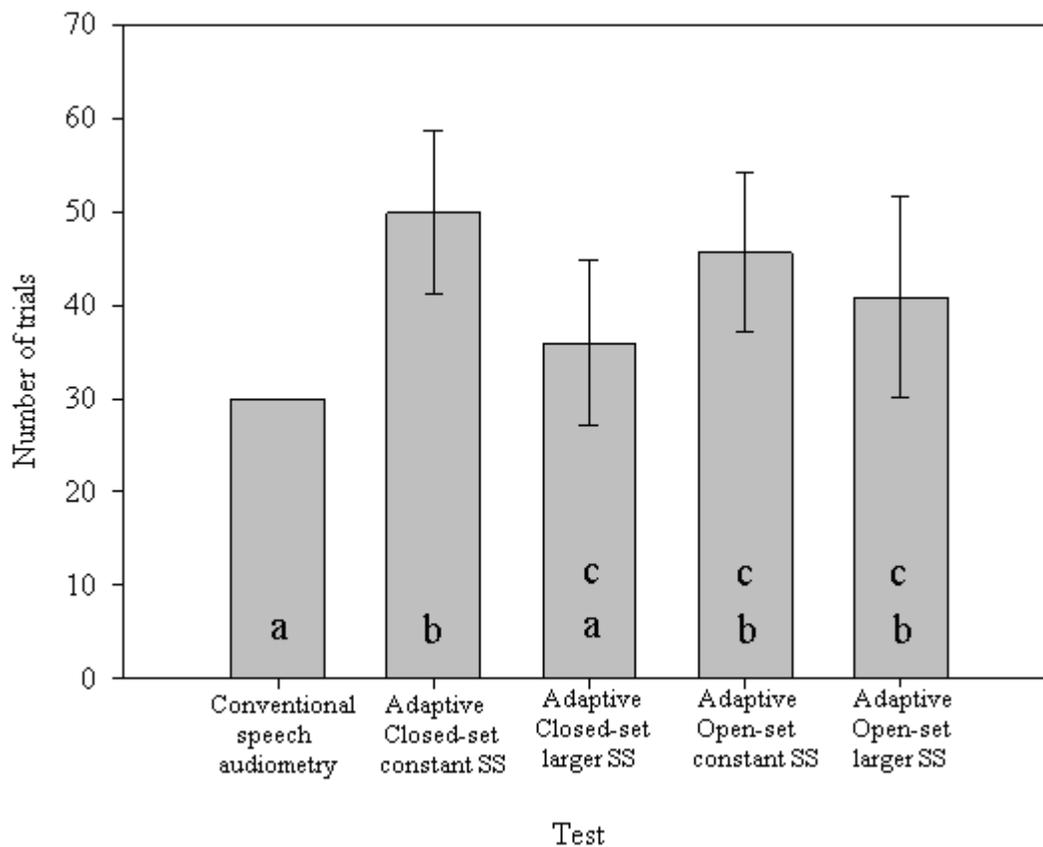


Figure 19. Mean number of trials required to reach a threshold within the 99% CI of the final threshold estimate on the adaptive tests. Group means that are significantly different from each other are labelled with different letters.

Results of a one-way RM ANOVA on Ranks revealed a significant test effect on the length of time required to reach a threshold within the 99% confidence interval of the final threshold estimate on the adaptive staircase tests (Chi-square = 34.520, $df = 4$, $p < 0.001$). As shown in Figure 20, the adaptive closed-set test with larger step sizes at the beginning required significantly less time to complete than conventional speech audiometry and both adaptive open-set tests ($p < 0.05$) while there was no significant difference in the time required to complete the two adaptive open-set tests, the adaptive closed-set test with constant step sizes and conventional speech audiometry.

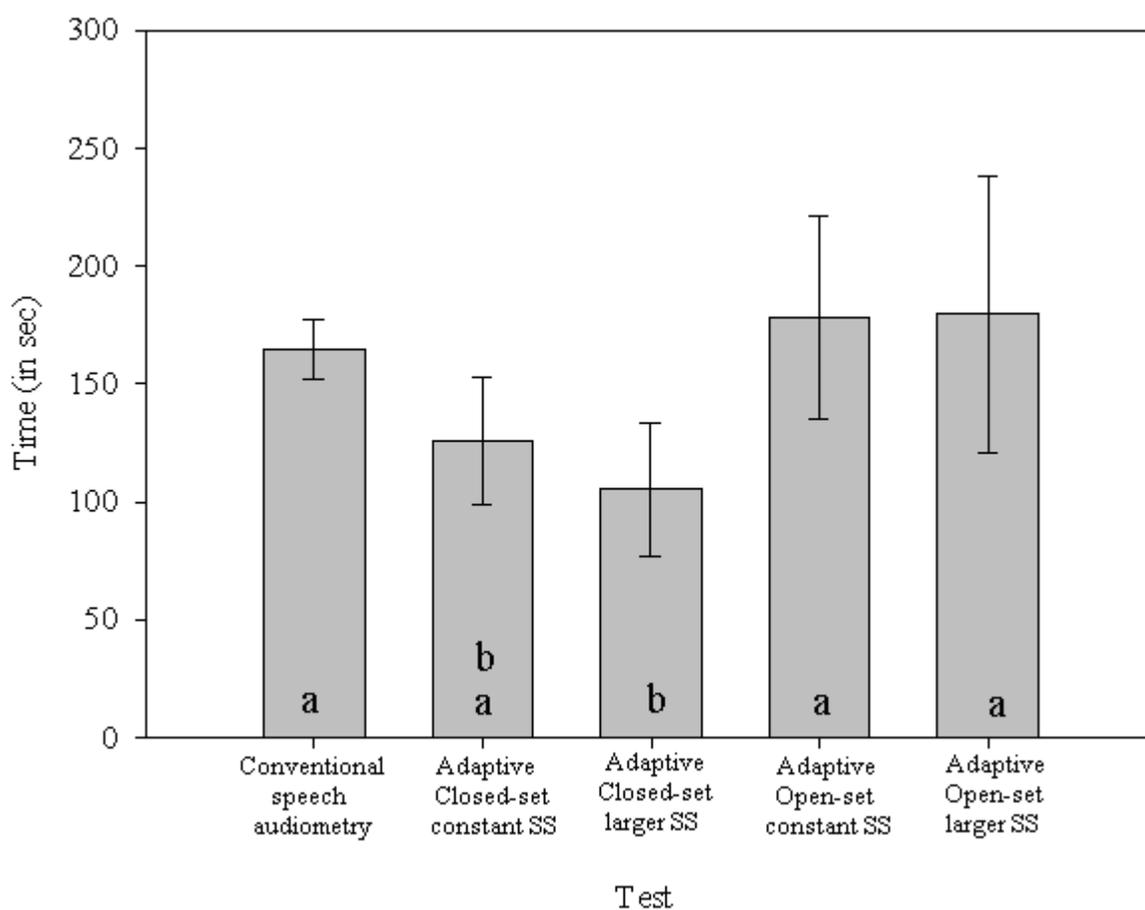


Figure 20. Mean length of time taken to reach the 99% CI of the final threshold estimate on the adaptive tests. Group means that are significantly different from each other are labelled with different letters.

Comparison of Termination Criteria

In order to ensure that the chosen termination criterion for use with each adaptive test allowed the majority (i.e. 95%) of the obtained threshold estimates to be within an adequate accuracy range of the final threshold estimate (as determined by the specific termination criterion), the average number of trials plus two standard deviations was used as an indication of the number of trials required for each adaptive staircase test used in the Clinical Testing Phase. As shown in Table 2, the termination criterion that allowed the least number of trials for the closed-set test with constant step sizes was the +4/-1 dB rule; that for the closed-set test with larger step sizes at the beginning and the open-set test with constant step sizes was the +3/-2 dB rule; and that for the open-set test with larger step sizes at the beginning was the ± 2.5 dB rule.

Table 2. Number of trials required for the threshold of 95% of participants to be within the accuracy range specified by the termination criterion employed in the adaptive staircase speech tests.

	Termination Criterion			
	± 2.5 dB	+3/-2 dB	+4/-1 dB	99% CI
Closed-set - constant SS	63	61	54	68
Closed-set - larger SS	54	52	54	54
Open-set - constant SS	51	50	60	63
Open-set - larger SS	48	49	59	63

Calculations based on the mean number of trials plus 2 standard deviations
All figures rounded up to the nearest trial

Individual Trial Completion Time:

Results of a one-way RM ANOVA revealed a significant test effect on individual trial completion time (Chi-square = 76.4, $df = 4$, $p < 0.001$). Results of post-hoc pairwise multiple comparison procedures using the Tukey test revealed that the length of time required for the completion of one trial for both adaptive tests using the closed-set response format and the adaptive open-set test with constant step sizes was significantly shorter than that of conventional speech audiometry (see Figure 21).

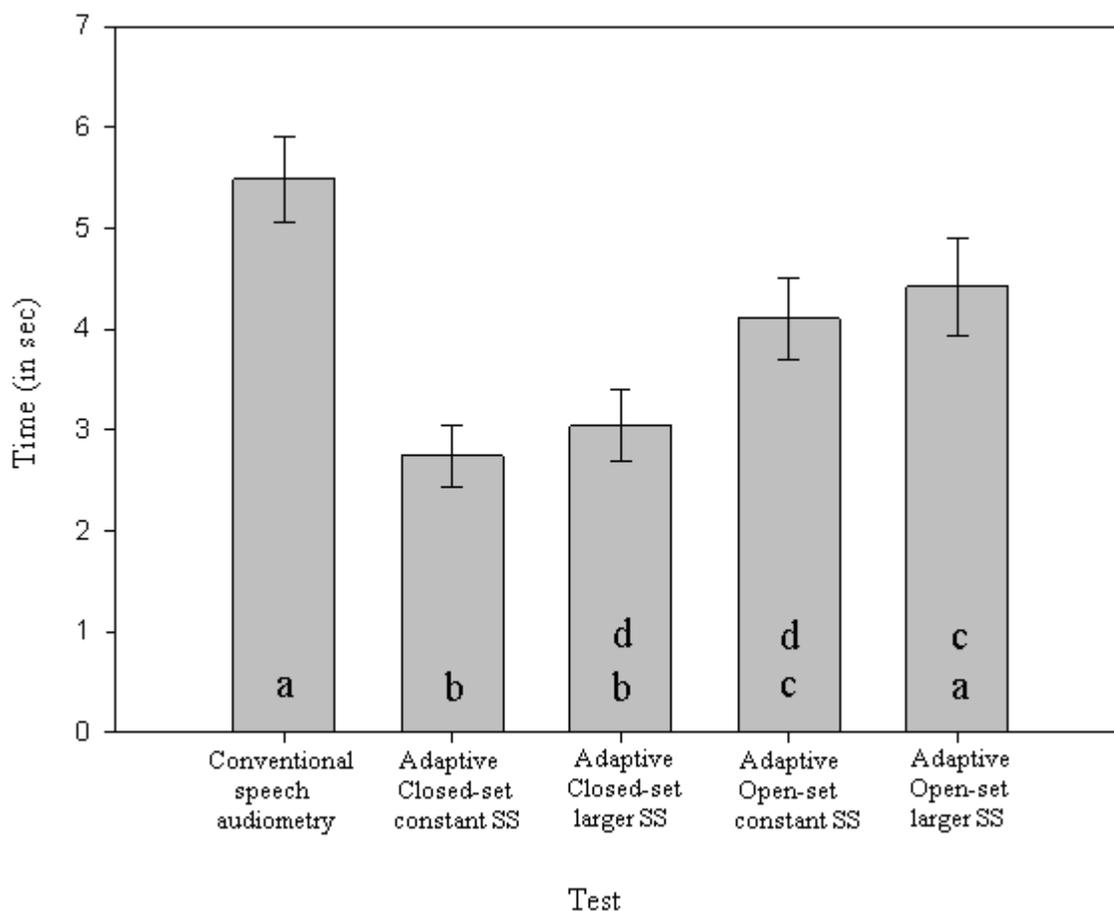


Figure 21. Mean length of time required for the completion of one trial (taken from the administration of one target word to the response of the participant to that target word). Group means that are significantly different from each other are labelled with different letters.

Step Size Variations - Constant Step Sizes versus Larger Step Sizes at the Beginning:

Results of a two-way RM ANOVA revealed a significant step size effect [$F(1,19) = 147.514, p < 0.001$], but no significant response format effect [$F(1,19) = 1.427, p = 0.247$] or step size by response format interaction effect [$F(1,19) = 4.108, p = 0.057$]. As shown in Figure 22, the constant step size procedures required a significantly larger number of trials to reach the same level of accuracy as that reached by the completion of the larger step size reversals in the procedures with larger step sizes at the beginning.

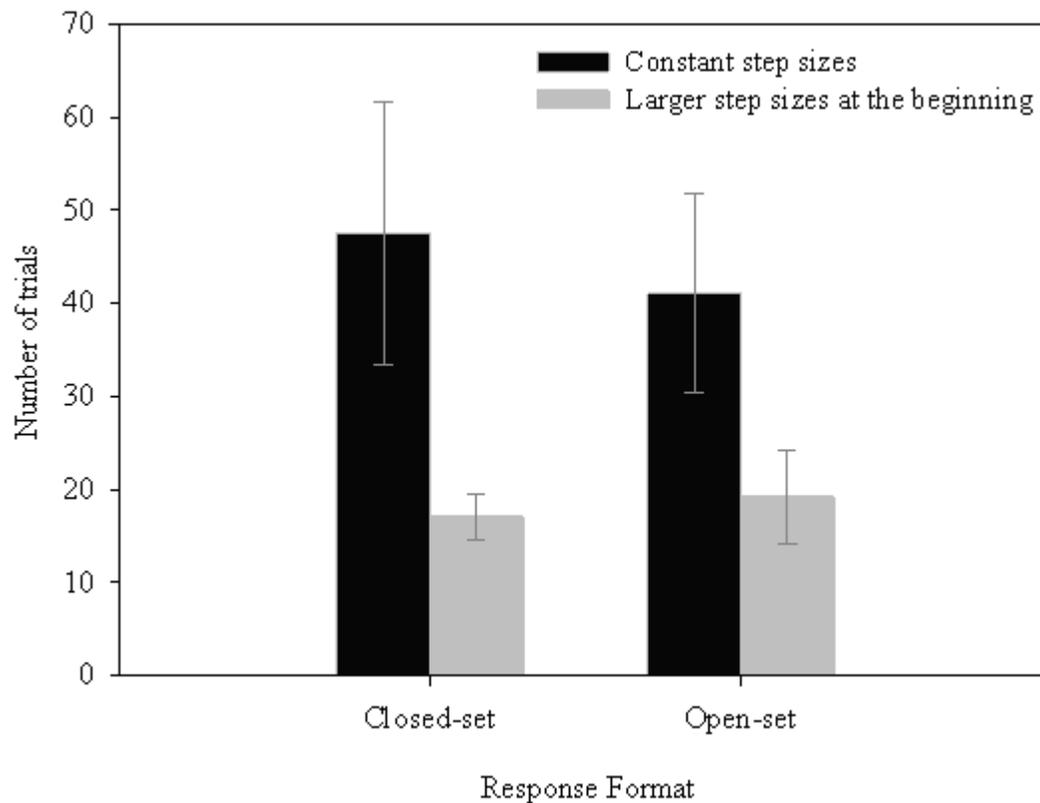


Figure 22. Comparison between the numbers of trials required for the adaptive procedures with constant step sizes to reach the same level of accuracy as that reached by the completion of the larger step size reversals in the equivalent procedures with larger step sizes at the beginning.

Results of a two-way RM ANOVA revealed a significant step size effect [$F(1,19) = 108.760, p < 0.001$] and response format effect [$F(1,19) = 22.408, p < 0.001$] but no significant step size by response format interaction effect [$F(1,19) = 0.0519, p = 0.822$]. As shown in Figure 23, the constant step size procedures required a significantly longer amount of time to reach the same level of accuracy as that reached by the completion of the larger step size reversals in the procedures with larger step sizes at the beginning. Additionally, the open-set tests required a significantly longer amount of time to complete than the closed-set tests.

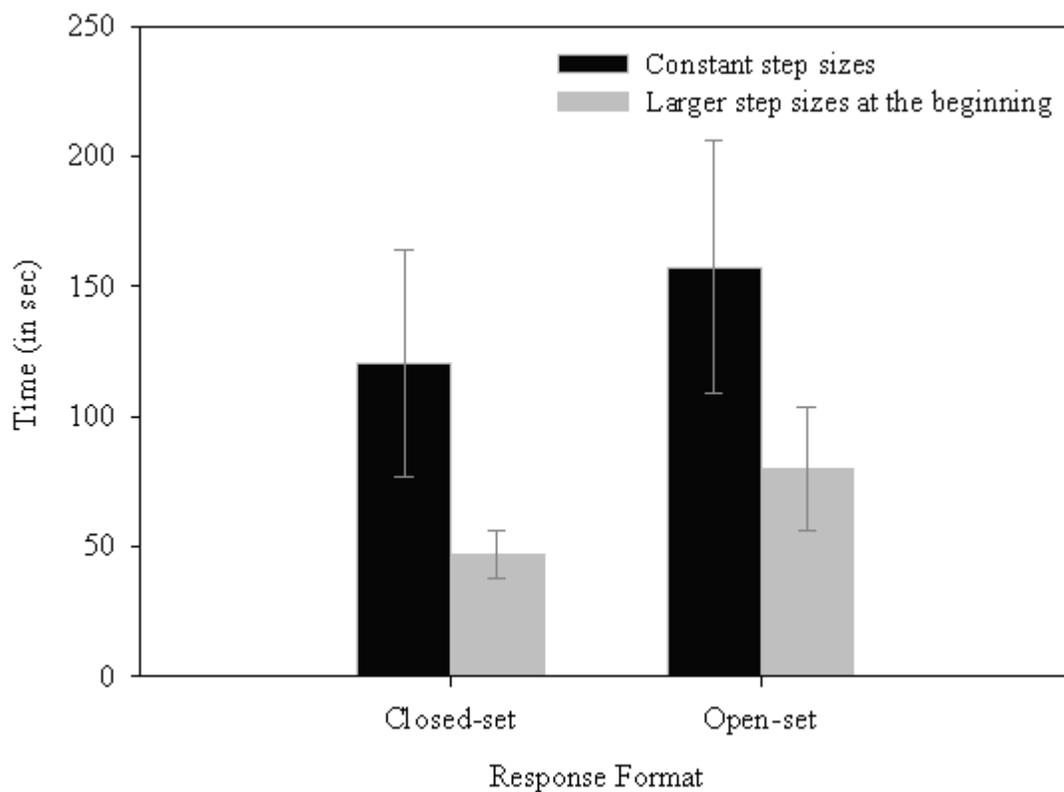


Figure 23. Comparison between the lengths of time required for the adaptive procedures with constant step sizes to reach the same level of accuracy as that reached by the completion of the larger step size reversals in the equivalent procedures with larger step sizes at the beginning.

4.3.1.3 Administration of Testing Procedure – Maximum-Likelihood (QUEST) Speech Tests

Results of a one-way RM ANOVA on Ranks revealed a significant test effect on the number of trials required to complete each adaptive maximum-likelihood speech test when using the ‘within ± 2.5 dB of the final QUEST estimate’ as a termination criterion (Chi-square = 38.160, $df = 3$, $p < 0.001$). In comparison with conventional speech audiometry, only the independent QUEST procedure required a significantly fewer number of trials for completion, whereas the adaptive staircase procedure with constant step sizes required a significantly larger number of trials for completion (see Figure 24).

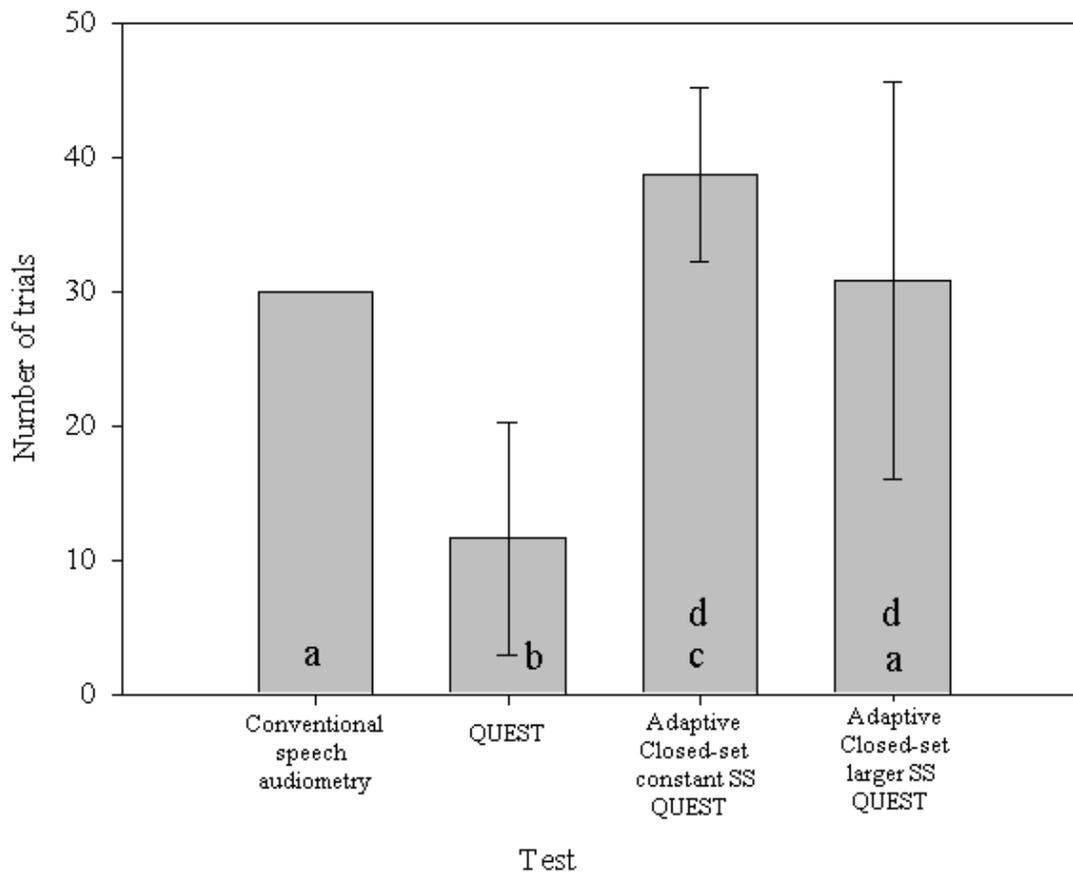


Figure 24. Mean number of trials required to reach within ± 2.5 dB of the final threshold estimate on the adaptive maximum-likelihood (QUEST) tests. The mean number of trials required to complete conventional speech audiometry is included as a reference. Group means that are significantly different from each other are labelled with different letters.

Results of a one-way RM ANOVA on Ranks revealed a significant test effect on the length of time required to complete each adaptive maximum-likelihood speech test when using the ‘within ± 2.5 dB of the final QUEST estimate’ as a termination criterion (Chi-square = 40.020, $df = 3$, $p < 0.001$). In comparison with conventional speech audiometry, all the adaptive tests required a significantly shorter length of time for test completion (see Figure 25).

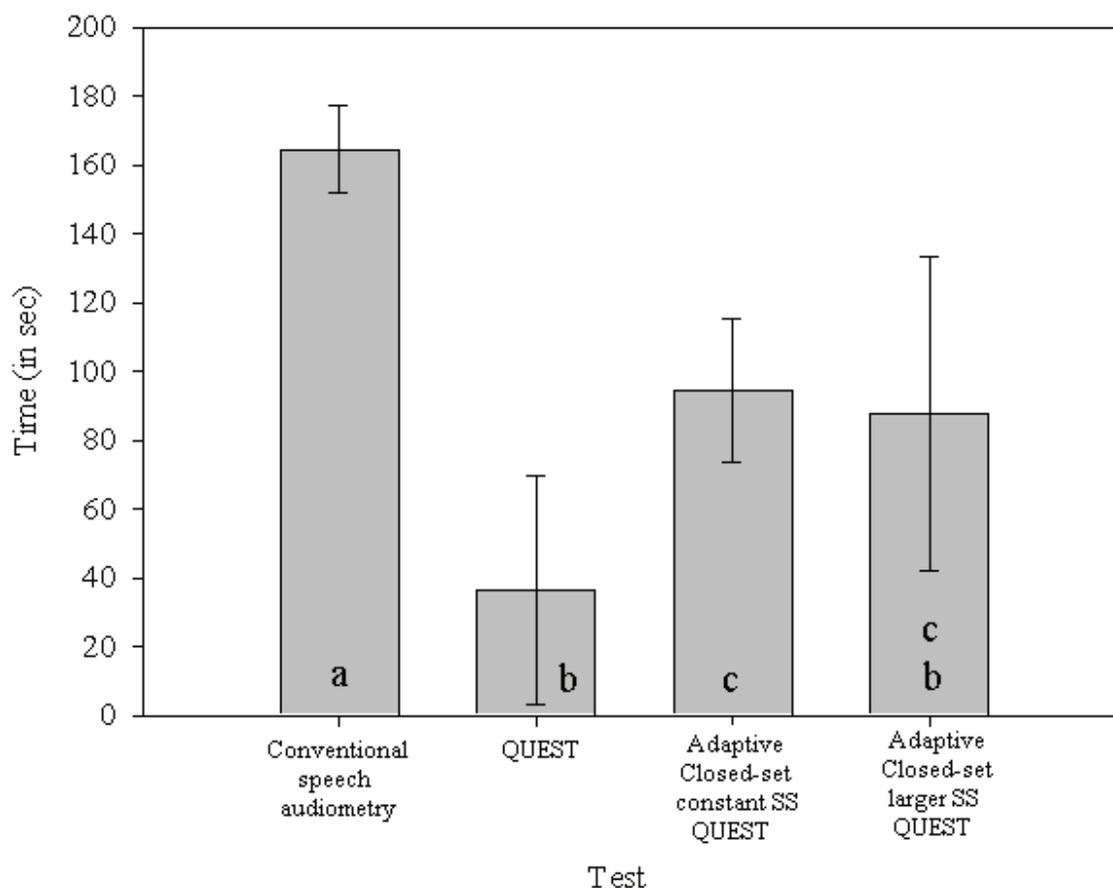


Figure 25. Mean time taken to reach within ± 2.5 dB of the final threshold estimate on the adaptive maximum-likelihood (QUEST) tests. The mean length of time taken to complete conventional speech audiometry is included as a reference. Group means that are significantly different from each other are labelled with different letters.

Termination Criteria and Number of Trials

In order to ensure that the chosen termination criterion allowed the majority (i.e. 95%) of the obtained threshold estimates to be within ± 2.5 dB of the final threshold estimate, the average number of trials plus two standard deviations was used as an indication of the number of trials required for each adaptive test used in the Clinical Testing Phase. As shown in Table 3, the independent QUEST procedure required the least number of trials, while the adaptive closed-set test with larger step sizes at the beginning required the largest number of trials.

Table 3. Number of trials required for the threshold of 95% of participants to be within ± 2.5 dB of the final QUEST estimate.

Procedure	Number of Trials Required
QUEST	29
Closed-set – constant SS QUEST	52
Closed-set – larger SS QUEST	61

Calculations based on the mean number of trials plus 2 standard deviations
All figures rounded up to the nearest trial

4.3.2 Accuracy

Moderate to high correlations are evident between the final SRT estimates obtained on all speech tests (Table 4). Higher correlations are evident between the SRT estimates from conventional speech audiometry and the open-set adaptive speech tests, when compared to those of conventional speech audiometry and the closed-set adaptive speech tests. When compared to the PTA, the closed-set test with constant step sizes and the open-set test with larger step sizes at the beginning show higher correlations compared to their equivalent procedures employing the alternative step size variation.

Table 4. Between test correlations on the accuracy of the SRT (50% correct threshold on open-set tests; 62.5% on closed-set tests).

		PTA	Adaptive Closed-Set		Adaptive Open-Set	
			Constant SS	Larger SS	Constant SS	Larger SS
Estimate from conventional speech audiometry		0.737	0.793	0.815	0.933	0.902
PTA			0.703	0.490	0.658	0.757
Adaptive Closed-Set	Constant SS			0.703	0.784	0.871
	Larger SS				0.807	0.784
Adaptive Open-Set	Constant SS					0.881
	Larger SS					

All correlations are significant at the $p < 0.05$ level

Adaptive SRT thresholds are based on the final threshold estimates obtained after 22 reversals

Results of a one-way RM ANOVA on Ranks revealed a significant test effect on the mean SRT estimate (Chi-square = 54.964, $df = 5$, $p < 0.001$). All SRT estimates obtained with the adaptive staircase tests, except those obtained with the adaptive closed-set test with larger step sizes at the beginning, were significantly lower than those obtained with conventional speech audiometry ($p < 0.05$), but were not significantly different from the PTA.

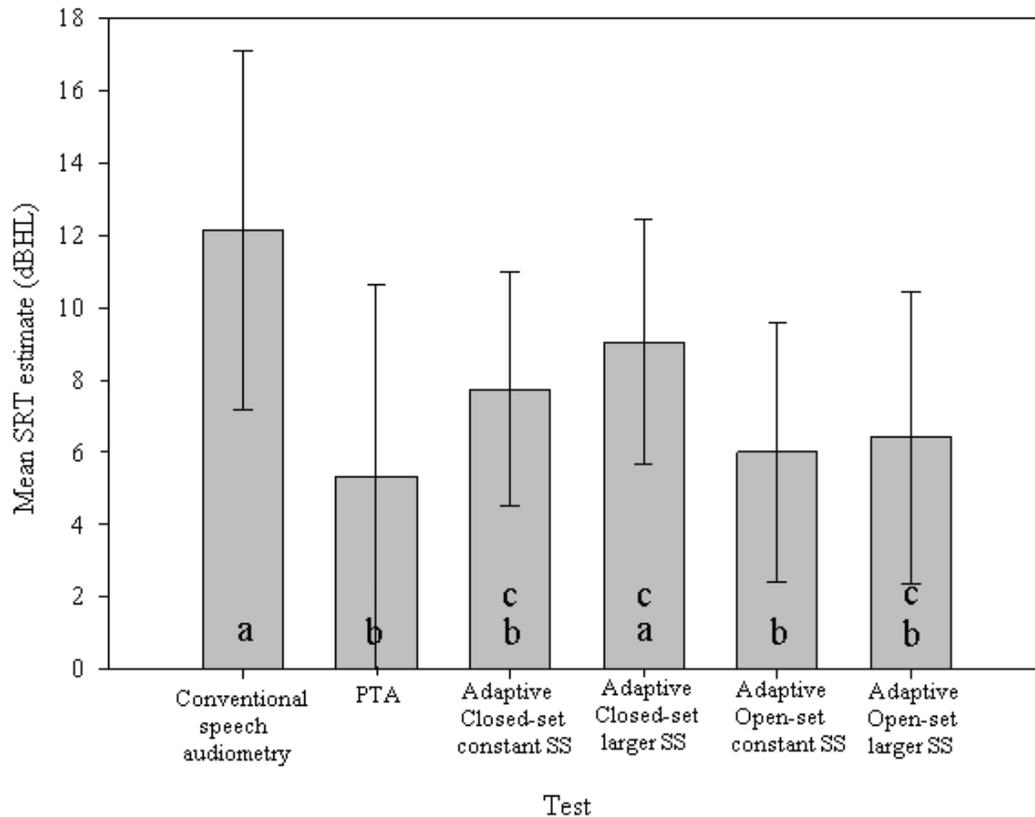


Figure 26. Mean SRT estimates obtained from conventional speech audiometry and the adaptive staircase procedures (The mean PTA value, a correlate of the SRT, is included for comparison).

When the level of the threshold estimate obtained from the adaptive closed-set constant step size procedure was used as the level for the administration of 20 words from the NU-CHIPS lists using the method of constant stimuli, the resulting percentage correct scores ranged from 35-75%, with a mean of 59.5% ($S.D. = 9.018$). Using this presentation level, 70% of participants scored within $\pm 7.5\%$ of the predicted 62.5% SRT score (i.e. between 55-70%), while only 30% scored within $\pm 2.5\%$ of their SRT value (i.e. 60-65%).

Moderate correlations are evident between the final 82%/76% threshold estimates obtained on all speech tests (Table 5).

Table 5. Between test correlations on the accuracy of the 82%/76% correct thresholds (82% on the closed-set tests; 76% on the open-set tests).

		Estimate from conventional speech audiometry	QUEST	Adaptive Closed-Set	
				Constant SS	Larger SS
Estimate from conventional speech audiometry			0.721	0.749	0.597
QUEST				0.659	0.670
Adaptive Closed-Set	Constant SS				0.608
	Larger SS				

All correlations are significant at the $p < 0.05$ level

Adaptive thresholds are based on the final threshold estimates obtained after 50 trials for the independent QUEST procedure, and after 22 reversals for the staircase procedures

Results of a one-way RM ANOVA revealed a significant test effect on the mean 82%/76% threshold estimate [$F(3,57) = 52.782, p < 0.001$]. As shown in Figure 27, all 82% threshold estimates obtained with the QUEST procedure and/or algorithm were significantly lower than the equivalent 76% threshold obtained with conventional speech audiometry. Additionally, the 82% thresholds obtained with the QUEST algorithm during the closed-set staircase procedures were significantly lower than those obtained with the independent QUEST procedure and conventional speech audiometry, but were not significantly different from each other.

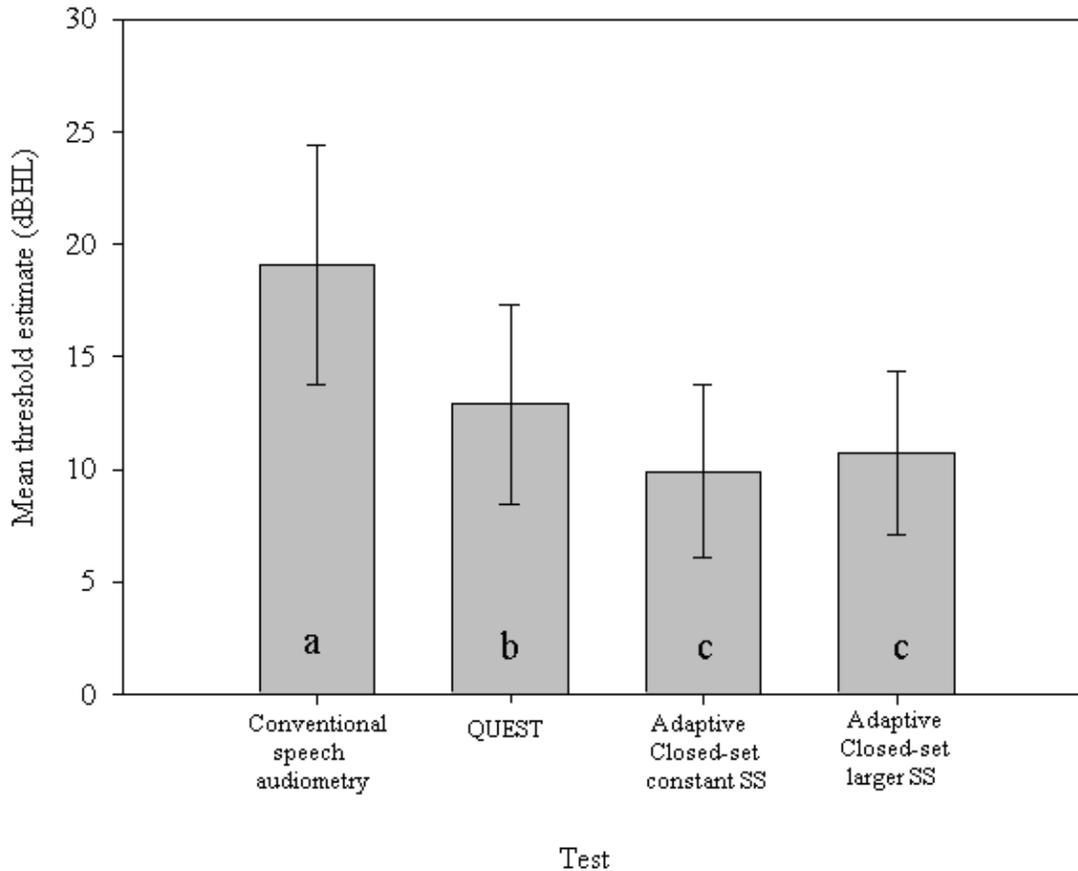


Figure 27. Mean 82% threshold estimates obtained with the adaptive tests, compared to the equivalent (76%) threshold estimates obtained with conventional speech audiometry. Group means that are significantly different from each other are labelled with different letters.

When the 82% threshold estimate obtained from the independent QUEST procedure was used as the level for the administration of 20 words from the NU-CHIPS lists using the method of constant stimuli, the resulting percentage correct scores ranged from 55-90%, with a mean of 78.75% (*S.D.* = 9.159). Using this presentation level, 85% of participants scored within +8/-7% of the predicted 82% QUEST score (i.e. between 75-90%), while only 30% scored within +3/-2% of their QUEST value (i.e. 80-85%).

4.3.3 Efficiency – Staircase Tests

The threshold estimates of 14 participants reached within ± 6 dB of their PTA on all four adaptive staircase tests (indicative of good agreement), while the threshold estimates of the remaining 6 participants reached within $\pm 7 - \pm 12$ dB of their PTA on all four adaptive staircase tests (indicative of fair agreement).

Results of a one-way RM ANOVA on Ranks revealed a significant test effect on efficiency when measured using the number of trials taken to reach a specified accuracy (Chi-square = 25.764, $df = 3$, $p < 0.001$). As shown in Figure 28, the adaptive staircase tests with larger step sizes at the beginning were more efficient than the equivalent adaptive tests with constant step sizes.

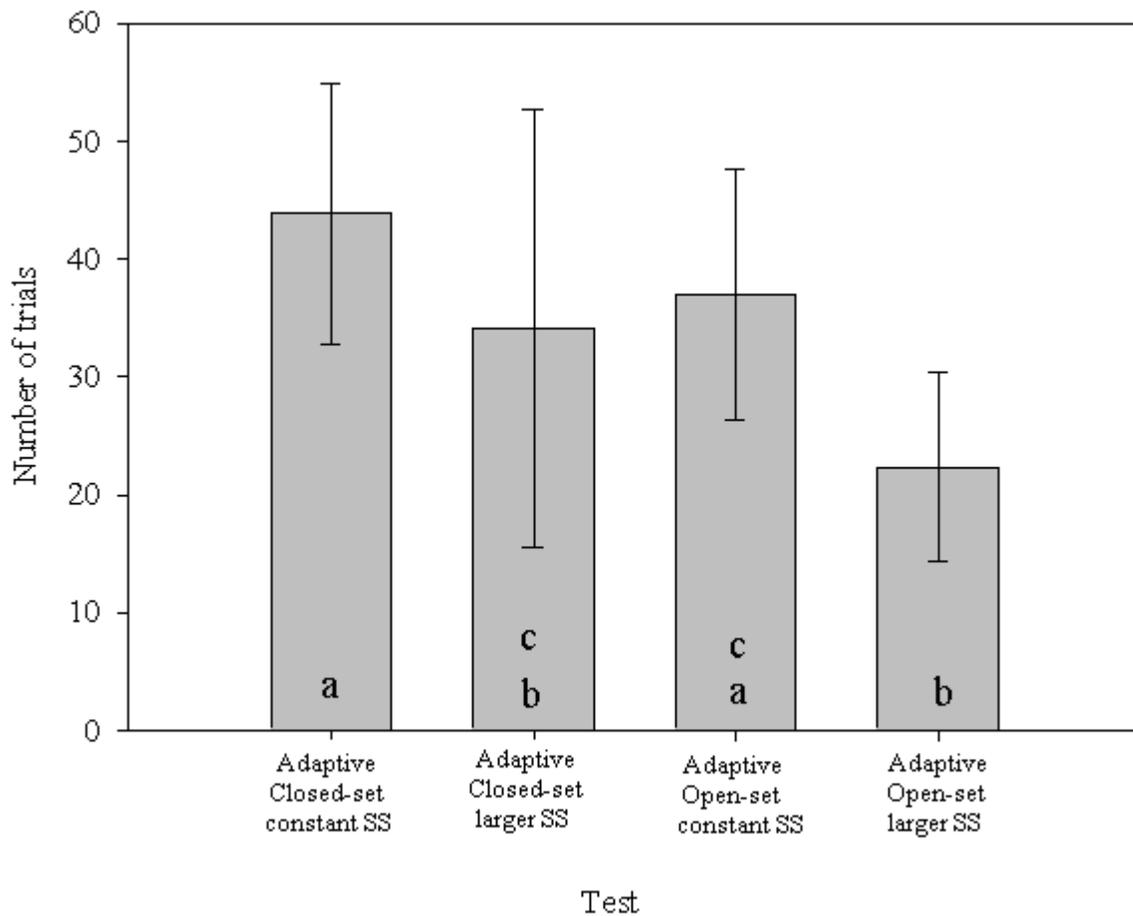


Figure 28. Mean number of trials required to reach the same accuracy range on all four adaptive staircase tests (i.e. the efficiency of each test) when using the PTA as an indicator of accuracy. Group means that are significantly different from each other are labelled with different letters.

Results of a one-way RM ANOVA on Ranks revealed a significant test effect on efficiency when measured using the time taken to reach a specified accuracy (Chi-square = 18.660, $df = 3$, $p < 0.001$). As shown in Figure 29, the adaptive open-set test with larger step sizes at the beginning was more time efficient than the equivalent test with constant step sizes; while there was no difference in the time efficiency between the two adaptive closed-set procedures.

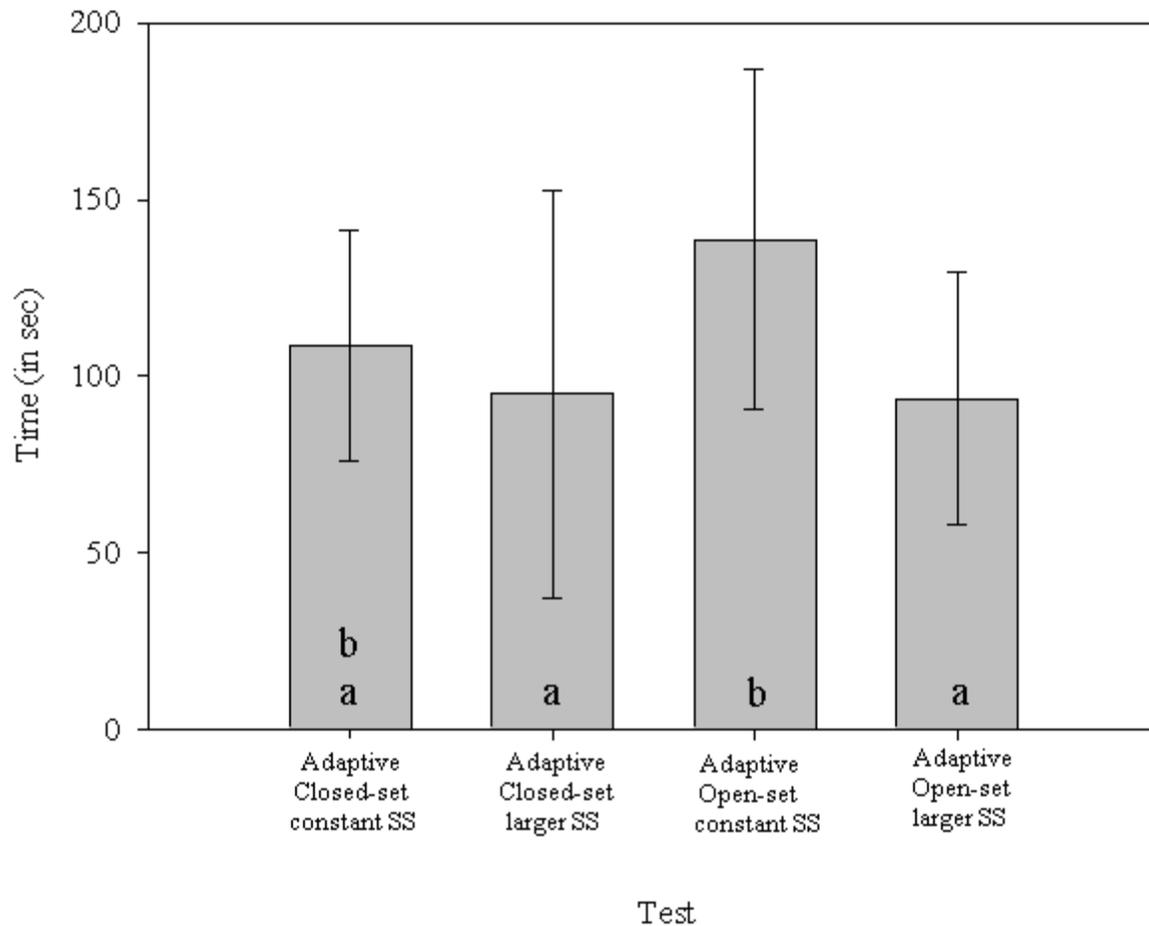


Figure 29. Mean length of time required to reach the same accuracy range on all four adaptive staircase tests (i.e. the efficiency of each test) when using the PTA as an indicator of accuracy. Group means that are significantly different from each other are labelled with different letters.

4.3.4 Reliability – Staircase and Maximum-Likelihood (QUEST) Tests

High correlations were evident between the first and second scores of the participants on both the closed-set and open-set adaptive staircase tests, with slightly higher correlations seen for the procedures with larger step sizes at the beginning (Table 6). Additionally, a strong correlation was evident between the first and second PTA values obtained at each testing session. The maximum-likelihood QUEST threshold estimates showed weaker correlations, with the only significant correlation evident between the first and second QUEST estimates from the adaptive closed-set test with larger step sizes at the beginning. No significant correlation was found between the first and second SRT scores obtained with conventional speech audiometry.

Table 6. Correlations between the first and second threshold estimates of 30% of the participants on each speech test, including the QUEST estimates based on the data from the closed-set staircase procedures.

Test	Correlation coefficient
Conventional speech audiometry SRT	0.753
PTA	0.935*
Adaptive Closed-set constant SS	0.877*
Adaptive Closed-set larger SS	0.978*
Adaptive Open-set constant SS	0.912*
Adaptive Open-set larger SS	0.919*
Conventional speech audiometry 76% threshold	0.849*
QUEST	0.645
QUEST estimate based on Closed-set constant SS test	0.667
QUEST estimate based on Closed-set larger SS test	0.823*

* Significant at $p < 0.05$ level

Paired t-tests and Wilcoxon Signed Rank tests revealed no significant differences (at the $p < 0.05$ level) between the first and second scores on any of the tests; however, since there were only six participants tested twice, the power of these statistical tests was low. The first and second SRT scores obtained using conventional speech audiometry differed on average

by 4.95 dB, while those obtained with the adaptive staircase tests differed by less than 2 dB on average. The 76% scores obtained with conventional speech audiometry differed by an average of 3.5 dB, while the equivalent 82% thresholds obtained with the independent QUEST procedure differed by 4.78 dB, and those based on the staircase procedures with constant and larger step sizes at the beginning differed on average by 3.76 dB and 2.52 dB respectively. There was no trend seen for the scores on any of the tests to increase or decrease on the second testing session, with both increases and decreases in thresholds being equally observed.

Examples of the independent QUEST tracking procedures for one participant (Participant Six) are displayed in Figure 30 to illustrate the difference in thresholds obtained during the first and second testing sessions. During the first testing session, the participant's early incorrect responses (trials 2, 6 and 7) caused the presentation level to remain at a relatively high intensity, while during the second testing session, the participant's early correct responses (trials 2, 4 and 6) caused the presentation level to remain at a relatively lower intensity. From the eighth trial onwards, neither procedure changed the presentation level more than 3.12 dB (during the first session it changed by 3.12 dB; during the second it changed by 2.39 dB), even though there was a considerable difference in the number of correct and incorrect responses that were made in these remaining trials during the two sessions - with only 3 incorrect responses in the final 42 trials in the first testing session (which equates to 92.86% correct), but 11 incorrect responses in the final 42 trials in the second testing session (which equates to 73.81% correct).

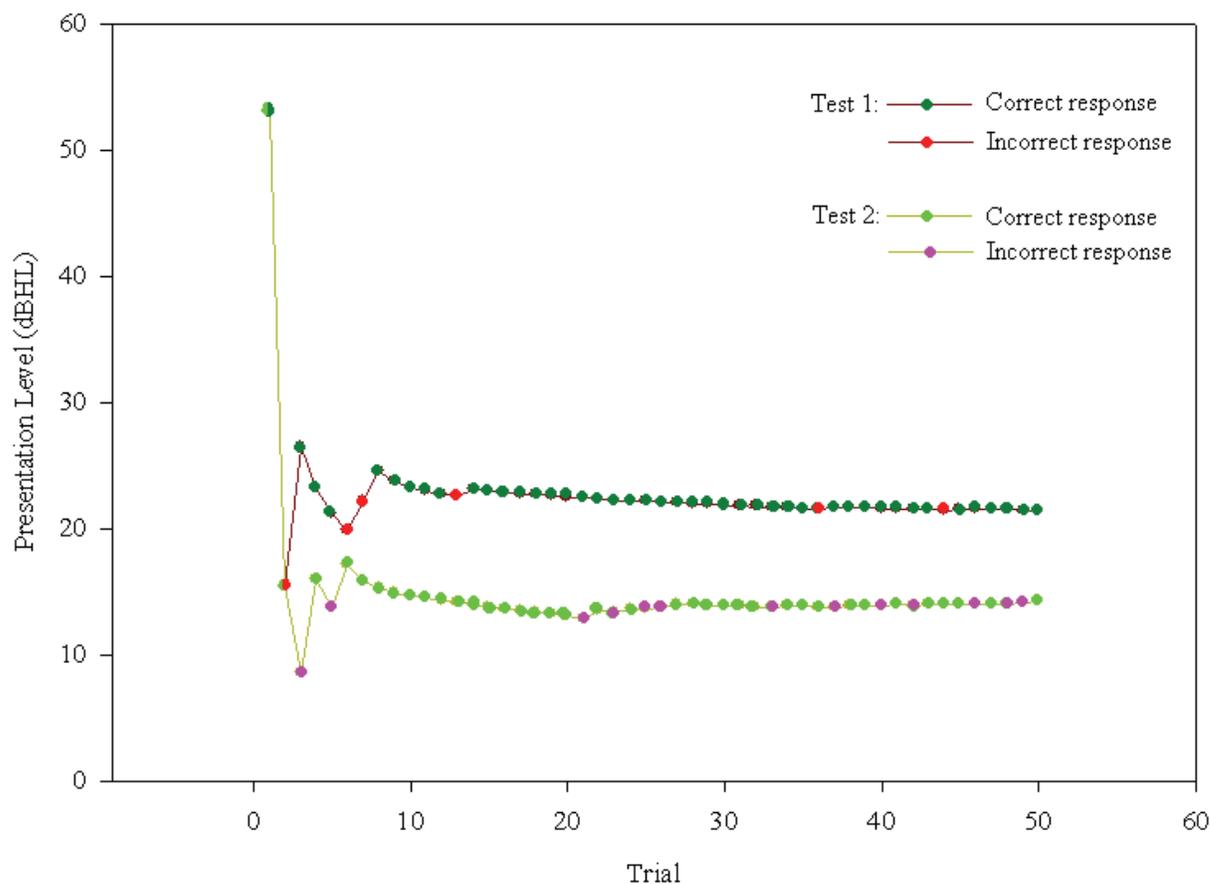


Figure 30. QUEST tracking procedures for Participant Six obtained during the first and second testing sessions.

4.4 QUEST Slope Estimates - 4AFC Tasks

Pooled data from the closed-set constant step size procedures of six normal hearing participants revealed that the slope of the resultant psychometric functions was 4.75 ($S.D. = 2.74$). This value is much higher than the presumed slope of 0.84 which was used in the implementation of the QUEST procedure in this study, which would have affected the degree to which QUEST could make good fits to the data and extract accurate thresholds.

4.5 Summary of Main Findings

The findings presented here are concerned with the determination of the optimal closed-set and open-set staircase procedures and the optimal maximum-likelihood procedure to be used in the subsequent Clinical Testing Phase. As such, the following sections deal exclusively with independent comparisons between the factors within each category of

adaptive test – closed-set staircase, open-set staircase and maximum-likelihood test. More detailed comparisons between adaptive and non-adaptive tests, closed-set and open-set tests and staircase and maximum-likelihood procedures are presented in Chapter Six.

4.5.1 Closed-Set Staircase Procedures

There was no clear optimal adaptive closed-set speech test, as both the procedure with constant step sizes and the procedure with larger step sizes at the beginning performed similarly in terms of time administration, accuracy, efficiency and reliability.

When used in conjunction with a 22 reversal and 99% CI termination criterion, the test with constant step sizes required a larger number of trials for completion than the test with larger step sizes at the beginning; however this trial difference did not translate into a time difference, as both closed-set adaptive tests required the same length of time for completion. Again, when used in conjunction with the three ‘within 5 dB of the final threshold estimate’ termination criteria, the test with constant step sizes required a larger number of trials for completion than the test with larger step sizes at the beginning when using the equivalent termination criteria; however, this did not translate into a time difference between the two tests. For 95% of participants’ thresholds to reach the required accuracy range specified by the termination criterion employed, the adaptive closed-set test with constant step sizes coupled with the ‘within +4/-1 dB of the final threshold estimate’ termination criterion and the adaptive closed-set test with larger step sizes at the beginning coupled with the ‘within +3/-2 dB of the final threshold estimate’ termination criterion were optimal, requiring 54 and 52 trials respectively.

In terms of the time taken to complete each trial there was no difference found between the two adaptive closed-set tests; however, the test with larger step sizes at the beginning required fewer trials and less time to reach the accuracy range at the end of the initial step sizes than the test with constant step sizes. As this finding is similar to that found with the adaptive closed-set tests, it indicates that the test with larger step sizes at the beginning approaches the vicinity of the threshold more rapidly than the test with constant step sizes,

and therefore focuses a larger number of trials around the estimated threshold, possibly increasing the accuracy of the final threshold value.

In terms of accuracy, the threshold estimates from both adaptive closed-set tests showed moderately high correlations with the thresholds obtained from conventional speech audiometry; however the test with constant step sizes showed a higher correlation ($r = 0.815$) with the PTA than the test with larger step sizes at the beginning ($r = 0.490$). There was no difference between the average threshold estimates obtained with the two adaptive closed-set tests, although the threshold obtained with the adaptive closed-set test with constant step sizes was lower than that obtained with conventional audiometry. The adaptive closed-set test with larger step sizes at the beginning was more efficient than the equivalent test with constant step sizes in terms of the number of trials required to reach a threshold estimate within a specified accuracy range of the PTA; however this did not translate into a difference in the time efficiency of the two tests.

The test-retest reliability was high for both adaptive closed-set tests, with a higher correlation seen for the test with larger step sizes at the beginning. Interestingly, the test-retest reliability of the adaptive closed-set test with larger step sizes at the beginning was the highest among all the speech tests administered in this study.

4.5.2 Open-Set Staircase Procedures

The adaptive open-set test with larger step sizes at the beginning, coupled with a ‘within ± 2.5 dB of the final threshold estimate’ termination criterion appeared to be the optimal adaptive open-set speech test for use in a clinical setting. This test showed advantages over the adaptive open-set test with constant step sizes with regards to administration time, accuracy and efficiency, and no significant difference with regards to reliability.

Although the number of trials and length of time taken to administer the two open-set adaptive procedures did not differ when used in conjunction with a 22 reversal or 99% CI termination criterion, the use of both the ‘within ± 2.5 dB of the final threshold estimate’ and the ‘within $+3/-2$ dB of the final threshold estimate’ termination criteria showed that the test

with larger step sizes at the beginning required less time for completion than the equivalent test with constant step sizes. For 95% of participants to reach the required accuracy range specified by the particular termination criterion employed, the test with larger step sizes at the beginning showed time advantages over the test with constant step sizes, with the least number of trials (48) required for the test with larger step sizes at the beginning when used in conjunction with the ‘within ± 2.5 dB of the final threshold estimate’ termination criterion. In terms of the time taken to complete one trial, there was no difference evident between the two adaptive open-set tests; however, the test with larger step sizes at the beginning required fewer trials and less time to reach the accuracy range at the end of the initial step sizes than the test with constant step sizes. This finding is similar to that found with the adaptive closed-set tests.

In terms of accuracy, the threshold estimates from both adaptive open-set tests showed high correlations with the thresholds obtained from conventional speech audiometry; however the test with larger step sizes at the beginning showed a higher correlation ($r = 0.757$) with the PTA than the test with constant step sizes ($r = 0.658$). There was no difference between the average threshold estimates obtained with the two adaptive open-set tests and the PTA, although both were lower than that obtained with conventional speech audiometry. The adaptive open-set test with larger step sizes at the beginning was also more efficient than the equivalent test with constant step sizes, evidenced by the fact that it required fewer trials and less time to reach a threshold estimate within a specified accuracy range of the PTA, which remained consistent for each participant across all the tests.

The test-retest reliability was high for both adaptive open-set tests, with a slightly higher correlation seen for the test with larger step sizes at the beginning.

4.5.3 Maximum-Likelihood Procedure/Estimates

When coupled with a ‘within ± 2.5 dB of the final threshold estimate’ termination criterion, the independent QUEST procedure required the least number of trials for completion (29 trials) compared to the number required for an equivalently accurate QUEST estimate from the adaptive closed-set staircase procedures. In these staircase procedures, the

test with constant step sizes had an advantage over the test with larger step sizes at the beginning insofar as the optimal number of trials for the determination of the SRT (54 trials) was above that required for determination of an adequately accurate QUEST estimate (52 trials); this was not the case for the test with larger step sizes at the beginning, as this required a larger number of trials (61 trials) than that required for the determination of the SRT (52 trials).

The independent QUEST procedure was the least reliable adaptive procedure, showing considerable variability between thresholds obtained on different testing sessions. There seemed to be a critical trial or trials near the beginning of the adaptive procedure where the response(s) of the participant had a larger impact on the rest of the procedure and, therefore, the final threshold estimate. The QUEST estimates based on the adaptive closed-set staircase procedures were also less reliable than the other adaptive threshold estimates, with the closed-set test with larger step sizes at the beginning showing better reliability than the test with constant step sizes.

The slope parameter employed in the QUEST procedure ($\beta = 0.84$) which was based on open-set speech audiometry, was much less than that calculated for the closed-set 4AFC speech test used in this study ($\beta = 4.75$). The degree to which the slope estimate was incorrect, however, was not known until after the completion of the study, and therefore, was not altered for use in the Clinical Testing Phase.

4.6 Conclusions - Speech Tests Chosen for Clinical Testing Phase

The speech tests chosen for administration in the Clinical Testing Phase were:

- a) Adaptive closed-set test – constant step sizes (54 trial termination criterion – based on an accuracy of $\pm 4/-1$ dB of the final threshold estimate);
- b) Adaptive closed-set test – larger step sizes at the beginning (52 trial termination criterion – based on an accuracy of $\pm 3/-2$ dB of the final threshold estimate);
- c) Adaptive open-set test – larger step sizes at the beginning (48 trial termination criterion – based on an accuracy of ± 2.5 dB of the final threshold estimate);

- d) QUEST (29 trial termination criterion – based on an accuracy of ± 2.5 dB of then final threshold estimate).

Although the optimal termination criterion for each adaptive procedure was determined during this Preliminary Testing Phase based on the number of trials required to ensure the majority (95%) of participants' thresholds reached an adequate accuracy level, the 22 reversal termination criterion was chosen to be employed in the Clinical Testing Phase to allow retrospective analyses of results based on these optimal termination criteria (as stated above). This would allow an evaluation of the adequacy of the optimal termination criteria identified in this Preliminary Testing Phase when used on a population that was more representative of that encountered in a clinical setting.

It was also decided that both adaptive closed-set tests be used in the Clinical Testing Phase of this study, as there was no clear optimal closed-set adaptive procedure identified during this phase. Likewise, despite the rather poor results obtained from the independent QUEST procedure, it was also chosen for administration in the Clinical Testing Phase to gather more data that could be helpful in determining the reasons for its poor performance. As it did not contribute greatly to the testing time, the longer 50 trial termination criterion was chosen for use with the independent QUEST procedure, so that results could be analyzed after the optimal number of trials (29).

Chapter 5
**Clinical Testing Phase of the Optimal Adaptive
Procedures**

5. Clinical Testing Phase of the Optimal Adaptive Procedures

The aim of this phase of the study was to test the optimal adaptive speech tests identified in the Preliminary Testing Phase on a population representative of that which typically attends an Audiology clinic.

5.1 Methods

The General Methods described in Chapter Three relate directly to this phase of the study. In addition, the characteristics of the participants and the specific speech tests that were administered to each participant are set out in the following sections.

5.1.1 Participants

Forty-six participants with varying degrees of hearing were recruited from the University of Canterbury Speech and Hearing Clinic to take part in the Clinical Testing Phase of the study. The participants consisted of 19 males and 27 females, and ranged in age from 14-81 years with an average age of 64.61 years (*S.D.* = 14.26). Hearing sensitivity, represented by the pure-tone average (PTA) of hearing thresholds at 500, 1000, and 2000 Hz, ranged from 5.0 – 80.0 dB, with a mean PTA of 22.870 dB (*S.D.* = 13.511). The participants formed a representative sample of those who would typically attend an Audiology clinic.

5.1.2 Speech Tests

Five speech tests were administered unaided and monaurally to the better hearing ear of each participant. These speech tests comprised:

1. Non-Adaptive Open-Set Test (Standard Speech Audiometry);
2. Adaptive Open-Set Test – simple staircase rule - larger step sizes at the beginning of the test;
3. Adaptive Closed-Set Test – weighted staircase rule – constant step sizes throughout the test;
4. Adaptive Closed-Set Test – weighted staircase rule – larger step sizes at the beginning of the test;
5. Adaptive Closed-Set Test – Maximum-Likelihood – QUEST.

5.1.3 Special Testing Circumstances

The starting level for the adaptive procedures administered to Participant 35, who had a PTA of 80 dB and known recruitment, was manually reduced by the examiner to avoid discomfort and excessive distortion. As the UC MAST program automatically began the adaptive procedures at a level 40 dB above the participant's PTA, a reduced value of 50 dB was entered into the UC MAST program in order to achieve a starting level of 90 dB (which was the maximum calibrated output of the equipment).

5.2 Results

5.2.1 Administration Time – Testing Procedures

A one-way RM ANOVA revealed a significant test effect on administration time (Chi-square = 135.304, $df = 4$, $p < 0.001$). As shown in Figure 31, both closed-set adaptive tests and conventional speech audiometry took significantly less time to complete than the adaptive open-set test, while the QUEST procedure required the least amount of time to complete.

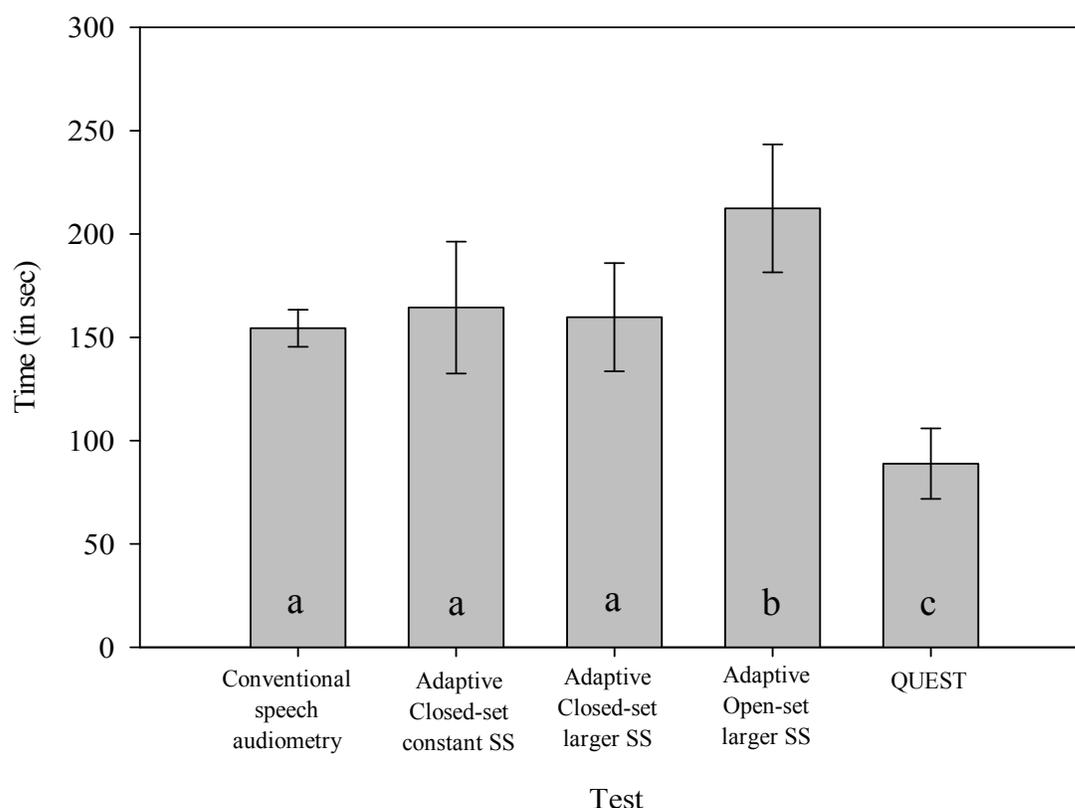


Figure 31. Mean length of time required to complete each speech test when using the least number of trials for each adaptive procedure based on the optimal termination criteria identified in the Preliminary Testing Phase (Closed-set constant step size: ‘within ± 4 dB of the final threshold estimate’ termination criterion, 54 trials; Closed-set larger step sizes at the beginning: ‘within ± 3 dB of the final threshold estimate’ termination criterion, 52 trials; Open-set larger step sizes at the beginning: ‘within ± 2.5 dB of the final threshold estimate’ termination criterion, 48 trials; QUEST: ‘within ± 2.5 dB of the final threshold estimate’ termination criterion, 29 trials). Group means that are significantly different from each other are labelled with different letters.

5.2.2 Accuracy

High correlations were evident between the SRT estimates obtained from conventional speech audiometry and all adaptive speech tests, as well as between all SRT estimates and the PTA (see Table 7).

Table 7: Between test correlations on the accuracy of the SRT (50% correct threshold on open-set tests; 62.5% on closed-set tests).

		PTA	Adaptive Closed-Set		Adaptive Open-Set
			Constant SS	Larger SS	Larger SS
Estimate from conventional speech audiometry		0.938	0.915	0.909	0.957
PTA			0.940	0.937	0.953
Adaptive Closed-Set	Constant SS			0.953	0.952
	Larger SS				0.935
Adaptive Open-Set	Larger SS				

All correlations are significant at the $p < 0.05$ level

Adaptive SRT thresholds are based on the threshold estimates using the optimal termination criteria and least number of trials identified in the Preliminary Testing Phase (Closed-set constant step size: 'within +4/-1 dB of the final threshold estimate' termination criterion, 54 trials; Closed-set larger step sizes at the beginning: 'within +3/-2 dB of the final threshold estimate' termination criterion, 52 trials; Open-set larger step sizes at the beginning: 'within ± 2.5 dB of the final threshold estimate' termination criterion, 48 trials).

Moderately high correlations were seen between the adaptive QUEST threshold estimates and the 76% threshold estimate obtained with conventional speech audiometry (Table 8). In all cases, a higher correlation was seen between the QUEST threshold estimate and the 76% threshold obtained from conventional speech audiometry when the QUEST estimates were derived from the staircase procedures rather than the independent QUEST procedure.

Table 8. Between test correlations on the accuracy of the 82%/76% correct thresholds (82% on the closed-set tests; 76% on the open-set tests).

		Estimate from conventional speech audiometry	Independent QUEST	Adaptive Closed-Set	
				Constant SS	Larger SS
Estimate from conventional speech audiometry			0.779	0.832	0.836
Independent QUEST				0.832	0.868
Adaptive Closed-Set	Constant SS				0.921
	Larger SS				

All correlations are significant at the $p < 0.05$ level

Adaptive thresholds are based on the threshold estimates using the ‘within ± 2.5 dB of the final threshold estimate’ termination criteria and least number of trials identified in the Preliminary Testing Phase (Closed-set constant step size QUEST: 52 trials; Closed-set larger step sizes at the beginning QUEST: 61 trials; independent QUEST: 29 trials).

Adequacy of Termination Criteria

The mean difference between the SRT estimate obtained after the optimal number of trials (identified in the Preliminary Testing Phase) and the final threshold estimate based on 22 reversals was smallest for the adaptive open-set test with larger step sizes at the beginning and largest for the adaptive closed-set test with constant step sizes (Table 9). The number of trials required to ensure that 95% of the participants obtained a threshold within the accuracy range specified by the optimal termination criterion identified in the Preliminary Testing Phase only held true for the adaptive open-set test, with over 95% of participants' thresholds reaching the specified level of accuracy. The adaptive closed-set test with larger step sizes came close to this level of accuracy, with just under 95% of participants' thresholds reaching the specified level of accuracy, while those of the closed-set test with constant step sizes and the maximum-likelihood QUEST procedures were much lower than predicted.

Table 9. The mean difference between the SRT thresholds obtained after the optimal number of trials identified in the Preliminary Testing Phase and the final threshold estimates obtained after 22 reversals, and the percentage of participants' thresholds within the accuracy range specified by the termination criterion employed in each adaptive test.

Adaptive Test	Difference between threshold after optimal number of trials and final threshold			Percentage of participants within the required accuracy range
	Mean (dB)	SD (dB)	Range (dB)	
Closed-set CSS	1.945	2.148	0 - 7.90	80.4%
Closed-set LSS	1.002	1.410	0 - 6.51	93.5%
Open-set LSS	0.680	0.693	0 - 3.65	97.8%
QUEST	1.698	1.932	0.06 - 10.75	78.3%
QUEST estimate from Closed-set CSS	1.039	1.524	0 - 8.01	91.3%
QUEST estimate from Closed-set LSS	0.230	0.409	0 - 2.01	100%

Final threshold is based on 22 reversals for the staircase procedures and 50 trials for the independent QUEST procedure

5.2.3 Efficiency

The threshold estimates of 22 participants reached within ± 6 dB of their PTA on all three adaptive staircase tests (indicative of good agreement), 22 reached within $\pm 7 - \pm 12$ dB of their PTA (indicative of fair agreement), while 2 did not reach within ± 13 dB (indicative of poor agreement).

The results of a one-way RM ANOVA on Ranks revealed a significant test effect on efficiency when measured using the number of trials taken to reach a specified accuracy based on the PTA (Chi-square = 23.374, $df = 2$, $p < 0.001$). The adaptive closed-set test and open-set test with larger step sizes at the beginning were both more efficient than the adaptive closed-set test with constant step sizes throughout (Figure 32).

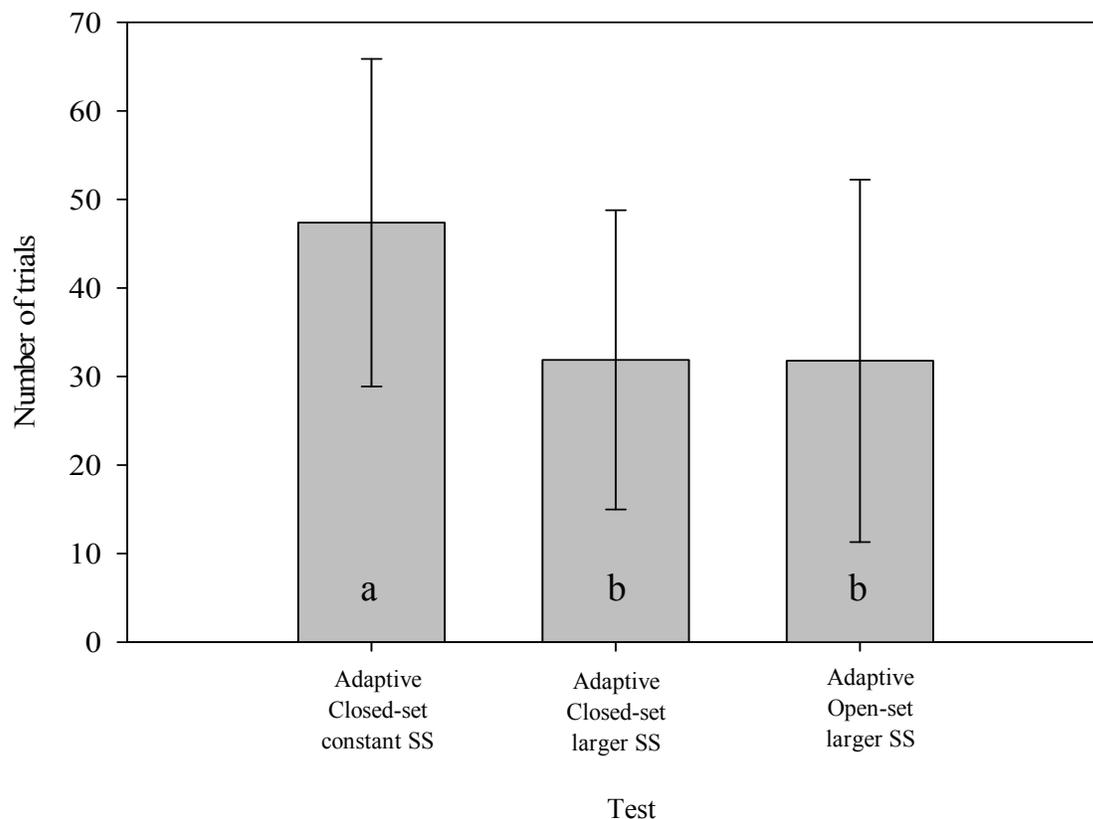


Figure 32. Mean number of trials required to reach the same accuracy range on all three adaptive staircase tests (i.e. the efficiency of each test) when using the PTA as an indicator of accuracy. Group means that are significantly different from each other are labelled with different letters.

The results of a one-way RM ANOVA on Ranks revealed a significant test effect on efficiency when measured using the time taken to reach a specified accuracy based on the PTA (Chi-square = 14.217, $df = 2$, $p < 0.001$). As shown in Figure 33, the adaptive closed-set test with larger step sizes at the beginning was the most efficient test; however, there was no significant difference between the time efficiency of the two tests with larger step sizes at the beginning (closed-set and open-set).

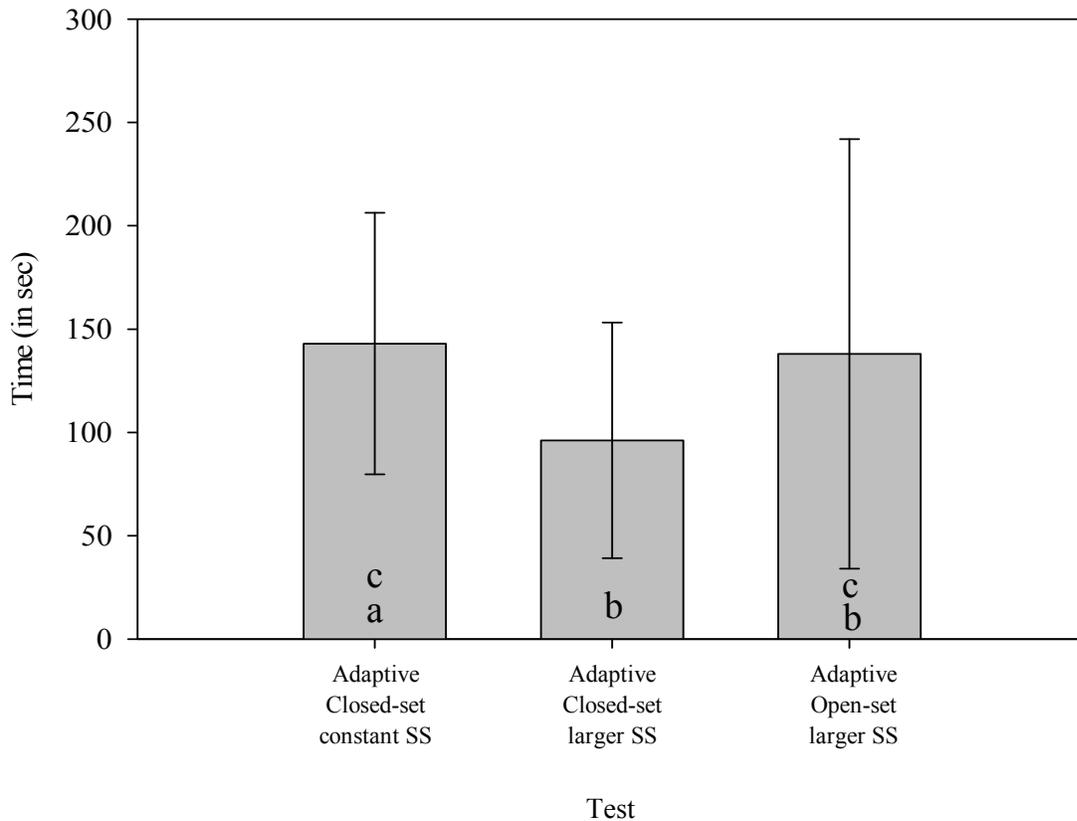


Figure 33. Mean length of time required to reach the same accuracy range on all three adaptive staircase tests (i.e. the efficiency of each test) when using the PTA as an indicator of accuracy. Group means that are significantly different from each other are labelled with different letters.

5.3 Summary of Main Findings

In comparison to conventional speech audiometry, only the independent QUEST procedure showed a significant time advantage, with both adaptive closed-set tests requiring the same length of time to complete, and the adaptive open-set test requiring a longer amount of time to complete. In terms of the accuracy of the threshold obtained after the optimal number of trials for each procedure, the increased length of time taken to administer the adaptive open-set test seemed to be advantageous, with >95% of participants' thresholds reaching the desired accuracy range of the final threshold estimate. The adaptive closed-set test with larger step sizes at the beginning showed an advantage over the equivalent test with constant step sizes in this respect, as a higher percentage of participants' thresholds reached the desired accuracy range of the final threshold estimate, suggesting that a preliminary phase with larger step sizes is beneficial in a clinical test. In keeping with this idea, the efficiency of both the adaptive closed-set and open-set test with larger step sizes at the beginning was higher than that of the adaptive closed-set test with constant step sizes throughout.

Chapter Six
General Discussion

6. General Discussion

The aim of the present study was to determine the optimal adaptive procedures and parameters for implementation into closed-set and open-set clinical speech tests, and assess the adequacy of the tests in terms of their administration time, accuracy, efficiency and reliability. Comparisons between the adaptive tests and conventional speech audiometry revealed the inadequacies of the current form of speech testing used in Audiology clinics, and highlighted the advantages and disadvantages of using different adaptive procedures as alternative means of testing speech discrimination abilities.

6.1 Comparison between Adaptive Methods and the Method of Constant Stimuli

6.1.1 Administration Time of Adaptive Procedures versus Conventional Speech Audiometry

When the adaptive speech tests were used in conjunction with the number of trials determined by the optimal termination criteria (in the Preliminary Testing Phase), almost all the adaptive closed-set tests (both staircase and maximum-likelihood procedures) required either the same amount of time or significantly less time to complete when compared to conventional speech audiometry. The only test which required a slightly longer length of time to complete (and only in the Clinical Testing Phase) was the open-set adaptive test with larger step sizes at the beginning. Although the speech tests did not use an equal number of trials, the individual trial completion time for each test revealed that both adaptive closed-set tests and the adaptive open-set test with constant step sizes required less time to complete each trial when compared to conventional speech audiometry, while the adaptive open-set test with larger step sizes at the beginning showed no difference. This finding is in keeping with the findings of Stach et al (1995), as the trial completion time for the majority of the adaptive tests (both closed-set tests and the open-set test with constant step sizes) employing a manual mode of presentation was shorter than that of conventional speech audiometry. This result also supports the first hypothesis of this study – that the adaptive tests would require less time to complete than conventional speech audiometry.

The manual mode of presentation of the adaptive tests was also advantageous insofar as it allowed the tests to adapt to both fast and slow responders, making them suitable for use with

a varied population. Additionally, this feature allowed the tests to adapt to the needs of the tester (e.g. if the clinician needed the opportunity to ask for clarification of an unclear response during the open-set adaptive tests) and also to the irregular time patterns that occur during atypical testing situations (e.g. test interruptions). Although this is a feature of all the adaptive tests used in this study, this feature is not specific to adaptive tests, and applies to all tests that require a response before the next stimulus is presented, i.e. any test with a manual mode of presentation.

Only the independent QUEST procedure showed a significant time advantage over conventional speech audiometry when compared using the optimal number of trials, partially consistent with the findings of Clark and Stewart (1968) and Turpin et al. (2002) which revealed time advantages of adaptive procedures in general. The lack of reliability of the QUEST procedure, however, and the inability to determine the accuracy due to the lack of a gold standard test (discussed in more detail in Section 6.6.1), meant that the time savings were not translated into increased test efficiency, and therefore were not clinically relevant.

6.1.2 Accuracy of Adaptive Procedures versus Conventional Speech Audiometry

There was a clear difference between the absolute thresholds obtained with the adaptive tests and conventional speech audiometry in the Preliminary Testing Phase, with those obtained with the adaptive procedures (both the SRT and 82% threshold) being lower, in keeping with the findings of Kollmeier et al. (1988), Saberi and Green (1997) and Dubno and Mills (1998). This finding may have been influenced by the increased number of trials placed in the vicinity of each participant's estimated threshold in the adaptive procedures, giving the participants more practice at responding to threshold stimuli, thus improving their scores.

The accuracy results lend partial support to the second hypothesis of this study - that the adaptive speech tests would show improved or at least comparable accuracy when compared to conventional speech audiometry (Buss et al., 2001; Taylor et al., 1983). Although there is no gold standard test with which to compare the results of the adaptive tests, the SRT values obtained with the staircase tests were consistently lower and closer to the PTA than those obtained with conventional speech audiometry. This consistency between the results of the

staircase tests and the PTA, coupled with their excellent test-retest reliability (which will be discussed in more detail in Section 6.1.4), indicates that the results do indeed possess a high degree of accuracy.

6.1.3 Efficiency of Adaptive Procedures versus Conventional Speech Audiometry

No direct comparisons were able to be made between the efficiency of the adaptive tests and that of conventional speech audiometry, due to the nature of the conventional speech audiometry test; therefore, the third hypothesis of this study – that adaptive procedures would be more efficient than conventional speech audiometry - was unable to be substantiated. Because the efficiency comparisons were based on the number of trials and time taken to reach a certain decibel level from each participant's PTA, conventional speech audiometry (in which the SRT is calculated as a fixed value only after three lists of words have been presented) was not able to be included in the comparisons. In keeping with the findings of Saberi and Green (1997) and Watson and Fitzhugh (1990), however, the administration time and accuracy data revealed that the adaptive staircase procedures seemed to be more efficient than conventional speech audiometry – taking the same or significantly less time to complete, and producing SRT values that showed high inter-test consistency and that were close to the PTA.

6.1.4 Reliability of Adaptive Procedures versus Conventional Speech Audiometry

Consistent with the views of Leek (2001) and Shelton et al. (1982), and in partial support of the fourth hypothesis, the adaptive tests in this study (excluding the maximum-likelihood QUEST procedures/estimates) produced more reliable threshold estimates than conventional speech audiometry. Whether this increased reliability was a by-product of an increased number of scorable items being used to estimate the adaptive thresholds as hypothesized, or whether it was an innate feature of the adaptive staircase procedures themselves, is unknown. No matter the reason for the increased reliability of the adaptive staircase tests, the results revealed the inadequate test-retest reliability of conventional speech audiometry as it is currently carried out in clinics. Repeated testing with the same participants using the current form of speech testing produced SRT values that were not significantly correlated and that differed by a large amount that could potentially produce clinically significant difference that

did not exist. This finding indicates that the clinical use of such a test to determine an individual's speech discrimination abilities, or their change in speech discrimination abilities over time, could produce results that do not reflect the actual abilities of the individual. The SRT values obtained with the adaptive staircase tests showed a higher correlation and differed by a much smaller decibel value, suggesting they are more suited for use in a clinical Audiology setting.

6.1.5 Additional Advantages of Adaptive Procedures

Speech tests employing the method of constant stimuli can be prone to 'ceiling effects' or 'floor effects', which can distort results and make it difficult to measure any improvement in speech perception abilities in certain populations (Helms et al., 2004). A 'ceiling effect' occurs when a number of participants score 100%, or close to 100% on a test, whereas a 'floor effect' occurs when a number of participants score 0%, or close to 0% (Mitchell & Jolley, 2004). Adaptive speech tests are flexible insofar as they adapt the presentation level of the stimuli to the abilities of the participant, and therefore help to avoid these ceiling and floor effects. Specific clinical situations in which the use of adaptive procedures would help reduce the occurrence of ceiling and floor effects include those situations in which one wants to assess long-term speech recognition improvement following some form of treatment, such as after cochlear implantation. In this situation, the use of current speech tests employing the method of constant stimuli are unsuitable, as many implanted individuals often reach close to maximum performance approximately one month after implantation, making further improvements difficult to track (Helms et al., 2004).

The adaptive speech tests used in this study also had advantages over conventional speech audiometry, as the clinician was not required to make any decisions or calculations prior to administration of the test/lists regarding where the patient's SRT was likely to occur. With conventional speech audiometry, the clinician chooses the levels at which to present the lists of words, introducing a variable which can affect the spread of the resultant scores, and therefore the ease at which a 50% threshold can be estimated. In the clinical setting, a multiple of 5 dB is usually chosen for testing (Martin, Champlin, & Chambers, 1998), and typically this does not give a percentage correct level of exactly 50%. In these cases the 50%

threshold has to be derived from the performance-intensity function, reducing the degree of accuracy as this is done by a visual best-fit, as opposed to employing statistical fitting procedures such as those used in this study.

6.1.6 Additional Disadvantages of Adaptive Procedures

Speech audiometry is a good indicator of a person's potential benefit from hearing aids, as it shows the result of amplification on speech discrimination abilities. With conventional speech audiometry, the first list that is presented aims to determine whether the person has any speech discrimination problems, by presenting speech stimuli at an audible level at which they should be able to get 100% correct. The adaptive procedures are unable to give this kind of information, because to track a 100% correct threshold is inherently non-adaptive, and would require the presentation of a certain number of words at a fixed intensity (i.e. a method of constant stimuli). Although the adaptive tests can be used to track a higher level on the psychometric curve (e.g. QUEST tracks the 82% correct threshold in a closed-set task), this is not high enough to rule out the possibility that the person suffers from some speech discrimination problems caused by distortion (such as that arising from impaired cochlear mechanics), whereby the clarity of the speech signal cannot be improved substantially regardless of increases in the presentation level of the stimuli (Carhart, 1951; Stephens, 1976). The kind of information obtained from determining whether a person is capable of achieving 100% on a speech discrimination task is useful in the clinical setting, as it shows the person's potential benefit from hearing aids and therefore contributes to the counselling of the patient with regard to realistic expectations with the use of hearing aids. This issue would be difficult to resolve clinically without the use of a non-adaptive method whereby a series of target stimuli are presented at a fixed supra-threshold intensity.

6.2 Comparison between Adaptive Procedures: Staircase Procedures versus Maximum-Likelihood Procedures

6.2.1 Administration Time of Staircase Procedures versus Maximum-Likelihood Procedures

As the tests were carried out on clinic patients, only the essential information required to carry out the tests was explained (and not the precise nature of the test and the means by

which the presentation level was controlled); thus the instructions for the staircase and maximum-likelihood closed-set tests were identical. No comparisons could therefore be made between the instruction administration time results found in this study and those found by Formby et al. (1996), as they compared the instructions for maximum-likelihood and staircase procedures. While a significant difference was found between the administration time of instructions between open-set and closed-set tests, it was on the order of 5 seconds and therefore would probably not result in a clinically important difference.

Contrary to the findings of Formby et al. (1996), the maximum-likelihood QUEST procedure took a significantly shorter length of time to administer than the adaptive staircase speech tests due to a reduced number of trials being required to reach a certain accuracy range of the final threshold estimate (determined after 22 reversals for the staircase tests and 50 trials for the QUEST procedure). This reduced administration time for the QUEST procedure, however, may have been due in part to the fact that 50 trials were not adequate for the determination of a suitably accurate threshold, and therefore the use of this final threshold value as a determinant of the accuracy was flawed. Additionally, the poor performance of the QUEST procedure in terms of the test-retest reliability highlighted the fact that the reduced administration time was only a byproduct of the fact that the accuracy of the final threshold estimate after 50 trials was not adequate, as will be discussed further in Section 6.2.2 below.

6.2.2 Accuracy of Staircase Procedures versus Maximum-Likelihood Procedures

The SRT values obtained from the different adaptive staircase procedures were highly correlated with each other and were consistent with the PTA, suggesting a high level of accuracy. In contrast, the 82% thresholds obtained using the QUEST procedure (and QUEST estimates based on the closed-set staircase procedures) showed weaker correlations with each other, and were consistently lower than the equivalent 76% estimates obtained with conventional speech audiometry. This finding, showing staircase procedures to produce more accurate results than maximum-likelihood procedures is inconsistent with the fifth hypothesis of this study and with the majority of previous research, which indicates that the two methods tend to produce results that are comparable in terms of accuracy (Buss et al., 2001; Formby et al., 1996; Gu & Green, 1994; Marvit et al., 2003). One possible explanation for the difference

between the results of this study and previous research is the fact that only one form of maximum-likelihood procedure (QUEST) was investigated, and the parameters used for its implementation were less than optimal as they were based on the use of a 2AFC task, as opposed to a 4AFC task. Additionally, the number of trials chosen for use with the QUEST procedure may not have been adequate to obtain the optimal accuracy. If this was the case, the QUEST procedure with the parameters employed in this study would not be suitable for clinical use, as the length of time required to administer more than 50 trials to find the 82% threshold would not result in the efficient use of clinical time.

The purpose of the seventh speech test used in the Preliminary Testing Phase of this study (comprising the administration of 20 words from the NUCHIPS word lists using a method of constant stimuli) was to crosscheck the accuracy of the thresholds obtained with the adaptive procedures. It was thought that presenting the words at the thresholds obtained using the adaptive procedures (closed-set constant step size staircase test and the independent QUEST procedure) would result in percentage correct scores that were close to those targeted by each procedure – i.e. 62.5% and 82% respectively. The results, however, revealed that the percentage correct scores obtained using this method of constant stimuli varied considerably from the thresholds targeted in the adaptive procedures. This larger degree of variance may have been due to the fact that only 20 words were administered, and therefore, each word contributed 5% to the overall percentage correct score. Any lapses in concentration on the part of the participant during this constant stimuli test would have affected the final score to a greater degree than during the adaptive tests, which comprised a larger number of stimuli. The validity of comparing the scores obtained from these procedures utilizing different numbers of trials was questionable, and therefore, the results of these comparisons were not taken to be representative of the adaptive procedures having poor accuracy.

6.2.3 Efficiency of Staircase Procedures versus Maximum-Likelihood Procedures

No comparisons were able to be made with regard to difference between the efficiency of the staircase procedures and maximum-likelihood QUEST procedure, predominantly due to the lack of knowledge regarding the true accuracy value of behavioural tests (Treutwein, 1995). Use of the sweat factor (Taylor & Creelman, 1967) as a measure of efficiency to

compare the two procedures would have been beneficial, as this statistic is able to summarize the efficiency of different procedures that target different thresholds and use different numbers of trials. Neither the sweat factor nor the method employed to calculate the efficiency of the staircase procedures were able to be employed successfully with the QUEST data, as there was no constant measure of accuracy with which to compare the 82% QUEST threshold. The equivalent 76% threshold obtained using conventional speech audiometry was not able to be used as a measure of accuracy due to the finding that this test had poor reliability, which brought into question the accuracy of the obtained threshold estimates.

6.2.4 Reliability of Staircase Procedures versus Maximum-Likelihood Procedures

Contrary to the fifth hypothesis of this study - that the staircase and maximum-likelihood procedures would produce results that were equally reliable - the staircase procedures were shown to produce more reliable threshold estimates. This finding is also contrary to the findings of Amitay et al. (2006) and Formby et al. (1996), who suggest both methods are equally reliable. Similar to the reasoning given for the difference in accuracy between the staircase and maximum-likelihood procedures used in this study, the reduced reliability of the QUEST procedure may be due in part to the inadequate number of trials used to obtain a threshold. The results of individual participants revealed the impact of early responses on the course of the adaptive track, and the relatively small intensity changes resulting from responses given later in the adaptive track. This observation indicates that the nature of the QUEST procedure may not be suitable for clinical use, as the information gained from the small intensity changes does not contribute to the overall threshold in a clinically significant manner – i.e. many of the intensity changes are fractions of a decibel, and therefore it takes a sequence of trials to initiate a clinically significant change in the final threshold estimate. This represents inefficient use of clinical time, and therefore, does not lend support for the use of the QUEST procedure (as it was employed in this study) as a clinical test.

6.3 Comparison between Closed-Set and Open-Set Adaptive Speech Tests

Advantages and disadvantages were evident with both the adaptive closed-set and open-set speech tests. In terms of administration time, the closed-set adaptive speech tests showed advantages over the open-set adaptive speech tests, with time savings evident during both the

administration of instructions and the testing procedures themselves. Although the absolute number of trials required to reach a 5 dB accuracy range of the final threshold estimate identified in the Preliminary Testing Phase was less for the optimal adaptive open-set test than the closed-set tests, the fact that each trial in the optimal adaptive open-set test required more time to complete led to the overall administration time difference. This finding is in keeping with the seventh hypothesis of this study - that the adaptive closed-set tests would require less time to complete than the adaptive open-set tests. The increased time required to complete each trial for the adaptive open-set tests was most likely due to the fact that they required an intermediary step in which the examiner had to manually score each response, whereas the adaptive closed-set tests were scored automatically by the computer program, thus saving time.

In addition to requiring a shorter administration time, the closed-set adaptive tests did not require any clinician time during the course of the testing procedure itself – as the scoring was carried out automatically by the computer program and did not require the clinician's attention. This automatic scoring system would provide benefits in a clinical setting, as it would increase time efficiency – e.g. the clinician could use the time while the patient was undergoing speech testing to carry out tasks that would otherwise have to be carried out later in the appointment, such as choosing suitable hearing aids for the patient's hearing loss or filling in appropriate hearing aid application/funding forms.

In the Preliminary Testing Phase, the accuracy of the thresholds obtained with the adaptive open-set tests tended to be lower than those obtained with the adaptive closed-set tests, and more consistent with the PTA. This finding is contrary to previous research, which has tended to show an improvement in the SRT when restrictions are imposed on participant responses, such as the use of closed-set response formats (Bruce, 1956). The comparisons made between SRTs obtained from the closed-set tests with those obtained from the open-set tests, however, were not entirely valid, as each response format utilized a difference scoring system – the open-set tests using a phonemic scoring system and the closed-set tests using a whole-word scoring system. Previous research has shown that there is a difference between the scores obtained with tests when using different scoring systems, with phonemic scoring yielding scores that are approximately 20% higher than those obtained from the use of

whole-word scoring (Olsen, Van Tassel, & Speaks, 1997). This simple 20% difference cannot simply be applied to the scores obtained from the tests used in this study, because the accuracy comparisons are confounded by the fact that, in addition to a different scoring system, the tests also used different response formats. Additionally, the lower thresholds obtained with the normal hearing participants on the adaptive open-set tests may have been due to a number of factors such as age and cognitive abilities. Research has shown that individuals with normal hearing often have the ability to fill in missing speech sounds even at barely audible levels (Eimas, Tajchman, Nyaaard, & Marcus, 1996), and young adults are more likely than older individuals to be able to predict words from context effects that take place at a phonological level, due to their increased cognitive abilities (Craig, Kim, Pecyna Rhyner, & Bowen Chirillo, 1993). As the individuals who participated in the Preliminary Testing Phase all had normal hearing and the average age was relatively low, these abilities, combined with a test utilizing a phonemic scoring system, would logically lead to improved scores when compared to those obtained on a test utilizing a whole-word scoring system.

The results of previous research showing that the accuracy of threshold estimates tend to increase as the number of choices in a forced-choice task increases (Amitay et al., 2006; Schlauch & Rose, 1990), could be extended with the use of open-set tasks, suggesting that the thresholds obtained with the open-set speech tests in this study are more accurate than those obtained with the closed-set tasks. Since the thresholds obtained with the adaptive open-set tests were consistently lower than conventional speech audiometry (as well as the closed-set adaptive tests), this suggests that the current form of speech testing overestimates participant's SRTs, and therefore may not provide an true representation of an individual's sensitivity for speech sounds.

The use of a phonemic rather than a whole-word scoring system gave the adaptive open-set tests an advantage over the adaptive closed-set tests. A phonemic scoring system gives a better indication of what speech sounds patients can hear, as many times they may be able to hear only a portion of the word – e.g. those with a high-frequency loss may be able to hear the vowel sound of the words, but may miss the softer, higher-pitched consonant sounds. Recent studies have supported this idea, with Amos and Humes (2007) showing that speech understanding is moderately and negatively correlated with the degree of high-frequency

hearing loss in elderly hearing-impaired listeners. By allowing the clinician to score the number of correct phonemes, the accuracy of the final threshold estimate obtained in the open-set tests is improved and the clinician has a clearer idea of the patient's reception of the acoustical cues of speech (Gelfand, 1998). Although unique to the open-set speech test used in this study (and not the closed-set test), a phonemic scoring system is not unique to open-set speech tests in general, and is a feature of some simple closed-set speech tests, for example the Minimal Pairs Test (Robbins, Renshaw, Miyamoto, Osberger, & Pope, 1988).

Although the SRT is determined by estimating the level at which a patient obtains 50% correct on the conventional open-set speech test, the assumption that chance performance was 0% in the adaptive open-set tests (in fact in any open-set speech test) is problematic. The comparison between the SRT estimates obtained from the adaptive open-set tests and conventional speech audiometry, however, did not pose any problems in this respect, as both procedures used an open-set response format, employed the same word lists coupled with a phonemic scoring system and targeted the 50% correct level. Comparing the two SRT estimates, therefore, was valid, even if neither targeted the exact 50% threshold. In theory, if a participant did not hear a single sound during the open-set speech test, it would be possible to score 0%; however, since each trial word comprises three scored components (phonemes), the audibility of one or two of these components would have an impact on the potential phonemes that made up the rest of the word. A common scenario for participants with normal low-frequency hearing thresholds but poorer high-frequency thresholds would be the audibility of the low-frequency vowel and consonant sounds but the inaudibility of the higher frequency consonant sounds. For example, for the target word 'cat', a patient may hear the 'c' and 'a' sounds, but fail to hear the 't' sound. In this example, the patient may still repeat the word 'cat', as there are limited possible phonemes that could complete the word – e.g. 'cat', 'cap' 'cad', 'can'. This restoration of the missing speech sounds from contextual information or the 'phonemic induction effect' (Warren, 1970) is a clear indication of the perceptual context effects on the SRT, and may have contributed to the lower thresholds obtained by participants on the open-set tests in the Preliminary Testing Phase.

While the closed-set adaptive procedure takes into account the fact that a person has a 25% chance of getting the word correct when they simply guess, there is a problem of partial

hearing. In cases where the participant hears part of the word, the probability of guessing the word correctly can increase – e.g. they may be able to eliminate one word from the alternative answers making their chance of getting the word right 33%, or may be able to narrow the correct answer down to two words, increasing their chance of correctly guessing the word to 50%. This problem is not taken into account in the closed-set adaptive tests; however, it is a problem innate to all closed-set tests and does not specifically apply to adaptive tests.

In general, there was no difference between the efficiency and reliability of the adaptive closed-set and open-set staircase speech tests. The efficiency of the optimal adaptive open-set test was equivalent to that of the best adaptive closed-set test (with larger step sizes at the beginning) in terms of the number of trials, and equivalent to both adaptive closed-set tests in terms of administration time. Reliability was high for all the adaptive staircase procedures, both closed-set and open-set.

6.4 Comparison between Step Size Variations in the Staircase Procedures: Constant versus Larger at the Beginning

In general, both the closed-set and open-set tests with larger step sizes at the beginning were more efficient than the equivalent tests with constant step sizes throughout. Consistent with the sixth hypothesis of this study, the tests with larger step sizes at the beginning reached the vicinity of the targeted threshold more quickly than the tests with constant step sizes throughout; however, contrary to this hypothesis, they also tended to require a fewer number of trials for completion (when using the ‘within 5 dB of the final threshold estimate’ termination criteria). These results are contrary to the recommendations given by García-Pérez (1998), who advised that it was more beneficial to expend the time that would have been used carrying out the preliminary larger step size phase performing a longer constant step size procedure. One possible reason for the disagreement between the results of the two studies is the fact that García-Pérez (1998) used simulations for visual detection tasks in an experimental type setting, and therefore could afford to use smaller step size variations as the length of time taken to administer the tests was not a significant issue. The fact that the current study used behavioural tests with speech discrimination tasks in a clinical setting

meant that a clinically relevant unit of change was required for the step sizes and the total administration time of the tests must be clinically acceptable.

6.5 Implications for Clinical Speech Audiometry

The findings of this study have a number of implications with regards to the way speech audiometry is currently carried out in the clinic. Although the time taken to administer conventional speech audiometry compared rather favourably with the time taken to administer the adaptive speech tests, inadequacies were highlighted in the test-retest reliability and accuracy (based on the absolute threshold values) of conventional speech audiometry. All the adaptive staircase SRT procedures were shown to have greater reliability than that of conventional speech audiometry, although due to the limited number of participants who were retested, more data would be required to substantiate this finding. The poor reliability of the SRT estimates obtained with conventional speech audiometry also brings into question the accuracy of the threshold estimates. In terms of the absolute threshold values obtained with each of the speech tests, the adaptive tests consistently produced lower estimates that were closer to the PTA than those obtained with conventional speech audiometry. The consistency between the thresholds obtained with the adaptive tests and their high correlations with the PTA, therefore, highlights the potential for conventional speech audiometry to overestimate an individual's 50% threshold, indicating an SRT that is poorer than the individual is actually capable of attaining.

6.6 Limitations of the Study

Although this study highlighted the advantages and disadvantages of adaptive procedures when implemented into clinical speech tests, there were a number of limitations in the design that may have influenced the results. As there was little previous research into the implementation of adaptive procedures into speech tasks, this study dealt with the complete development of 'optimal' adaptive speech tests, from the particular adaptive procedure used, to a number of more specific parameters such as step size variations and termination criteria. Due to the large nature of such an undertaking, focus was given to the key components of a speech test; therefore, it is inevitable that further modifications would need to be made to

these ‘optimal’ adaptive speech tests before they were suitable for use with *all* patients encountered in a clinical setting.

6.6.1 Validity of Accuracy Comparisons

The fact that the thresholds obtained from speech testing are based on the results of behavioural tests alone placed restrictions on the degree to which the threshold results from different speech tests could be compared; thus the interpretation of some of the results must be viewed with a degree of caution. The determination of the accuracy and efficiency of the adaptive speech tests was made difficult due to the fact that there was no gold standard test that determined with certainty an individual’s SRT, 76% or 82% threshold to provide a comparison to the adaptive threshold estimates. With the staircase procedures, the final threshold estimate obtained after 22 reversals was used to compare the accuracy of provisional estimates using different termination criteria, as previous studies have shown that the use of a much lower number of reversals was enough to obtain statistically significant results (e.g. Mackie & Dermody, 1986), and it was thought that after 22 reversals the threshold estimate should have stabilized to a point that was sufficiently accurate for clinical purposes. The results, however, showed that there was still considerable variability evident between the final threshold estimates of the same participants on different adaptive tests, and therefore the accuracy results obtained using different termination criteria were relative to the accuracy of the particular adaptive test. As there was no gold standard measure by which the most accurate threshold estimate could be determined, it was impossible to determine *with certainty* which speech tests were the most accurate.

One constant measure of accuracy, employed in the determination of the efficiency of the adaptive tests, was the degree of consistency between each participant’s SRT estimate and their PTA. Unlike the final threshold estimate obtained after 22 reversals (which differed between tests), the PTA of a participant remained constant. This consistency, coupled with the high test-retest reliability of the PTA, allowed a more valid comparison between tests to be made. One limitation with the use of the PTA as a comparison for accuracy, however, is the fact that there is debate regarding the exact correlation that can be expected between the

PTA and the SRT, and the thresholds that should be used to calculate the PTA (Carhart, 1946; Carhart & Porter, 1971; Fletcher, 1950). The three-frequency PTA (Carhart, 1946) employed for comparisons in this study is not always the combination of thresholds that is best correlated with the SRT, especially in cases of sloping hearing losses. Later research has indicated that the best two frequencies from 500, 1000 and 2000 Hz (Fletcher, 1950; Schlauch, Arnce, Olson, Sanchez, & Doyle, 1996), or subtracting 2 dB from the average of the thresholds at 500 and 1000 Hz (Carhart, 1971) results in better agreement with the SRT under some circumstances, especially when the patient has a sloping hearing loss. In cases of precipitously sloping hearing loss, it has been suggested that the single frequency that has the best threshold, as opposed to the average of several frequencies, is a better predictor of a person's SRT (Gelfand & Silman, 1985; Silman & Silverman, 1991). More recent studies have reiterated the continued lack of sound knowledge concerning the factors influencing the SRT (Picard, Banville, Barbarosie, & Manolache, 1999), with different studies using different combinations of pure-tone thresholds to make comparisons with the SRT (e.g. Marcell, 1995; Schlauch et al., 1996). Additionally, research has indicated that there are a number of other factors, apart from the PTA, that influence the SRT, most notably the steepness of the hearing loss (Carhart & Porter, 1971) and perceptual and cognitive linguistic factors such as the meaningfulness of the speech stimuli (Picard et al., 1999; Samuel, 1987). These additional influences were not factored into the comparisons made between the PTA and the SRT estimates, and therefore any discrepancies cannot be attributed with certainty to test inaccuracies. Although it was beyond the scope of the current study, future studies could take into account the differing research with regard to the PTA-SRT relationship by analysing the correlations between the SRT obtained with each test to both single frequencies and PTAs calculated using different methods.

6.6.2 Equipment Limitations

Limitations with the equipment used to administer the adaptive speech tests in this study were highlighted during the testing session involving the participant with a PTA of 80 dB HL (Participant 35 in the Clinical Testing Phase). Because the theoretical starting level of the adaptive tests for this participant was so high (40 dB above her PTA, i.e. 120 dB HL) the equipment, which could only be calibrated to a maximum of 90 dB HL was unable to

accommodate this intensity level successfully. To compensate for this problem, the participant's PTA was entered into the program at a reduced value of 50 dB HL, which resulted in a starting level of 90 dB HL. Even though the equipment was able to be calibrated up to this intensity level, there was significant distortion in the output of the laptop when target words were presented at 85 dB HL and above, most likely due to quality of the sound card. A 24-bit soundcard would have a dynamic range of 144 dB, rather than the 96 dB range of the 16-bit soundcard used in this study. Use of adaptive speech tests in a clinical setting would obviously require the use of good quality laptop or computer systems; however, as many clinics already use computer systems for the administration of pure-tone audiometry or hearing aid real-ear measures, such as the AVANT Stealth Audiometer and the AURICAL Plus, these would most likely be readily available for use.

6.6.3 UC MAST Program Limitations

Recruitment, an abnormal growth in loudness which often accompanies hearing loss, was not taken into account in any of the adaptive procedures in terms of the starting presentation level of the speech stimuli. In all cases, the adaptive tests began with a trial presented at 40 dB above the participant's PTA. In the case of the participant with a PTA of 80 dB HL (Participant 35 in the Clinical Testing Phase), who had known recruitment, a starting level of 120 dB HL would have been too high, and therefore would not have been appropriate even if the equipment was able to reach that intensity level without excessive distortion. As a result of this, more recent versions of the UC MAST program take the issue of recruitment into account, by allowing the clinician to have direct control over the starting level of the first trial (instead of having to enter the PTA in the program at a reduced level to obtain the desired starting level).

The UC MAST program did not employ the use of any contralateral masking noise, even when it was required according to standard speech audiometry protocols. The interaural attenuation for speech sounds when using insert earphones has been found to be between 55 - 80 dB, depending on the type of earphone and the depth of insertion (Goldstein & Newman, 1994). In current clinical protocols, the use of an interaural attenuation of 50 - 55 dB is commonly used to determine the need for masking. When the presentation level of the stimulus word to the test ear minus this interaural attenuation is better than or equal to the

best pure-tone bone conduction threshold in the non-test ear, masking is required. In most cases in this study, the first trials presented at 40 dB above the PTA of each participant would have crossed over to be heard in the non-test ear, and thus would have required contralateral masking to ensure the test-ear was being tested independently. In cases where there is a significant asymmetry between the hearing thresholds in each ear, masking is likely to be required when testing the poorer ear. In this study, this problem was avoided as only one ear was tested and it was always the better hearing ear in the Clinical Testing Phase where asymmetrical hearing losses could have posed a problem. Subsequent versions of the UC MAST adaptive speech test have therefore been modified to incorporate contralateral masking that is automatically applied when the presentation level of the stimulus word exceeds the intra-aural attenuation of the transducers and cross-hearing is a possibility. This requires each participant's audiogram to be pre-entered into the program in order to determine when contralateral masking is necessary, but results in more accurate testing of participants with large hearing threshold asymmetries.

6.7 Directions for Future Research

The limited previous research concerning adaptive open-set procedures, and the good performance of the adaptive open-set speech tests used in this study, implies that future research into this area could only be beneficial. The use of adaptive procedures with open-set tasks with varied scoring systems would open up a wide range of additional uses for adaptive procedures, and possibly widen their use in both experimental and clinical settings.

Further research concerning the optimal parameters for the implementation of an adaptive maximum-likelihood QUEST procedure into a clinical speech test with a 4AFC response format is also required because of the relatively poor performance of this procedure in this study. It would be interesting to compare the performance of the QUEST procedure targeting the 72% correct threshold (as this was determined to be the most efficient target for a QUEST procedure coupled with a 4AFC task) with the QUEST procedure used in this study targeting the 82% correct threshold. Increasing the number of trials used in the QUEST procedure would also be beneficial to determine whether this would increase the accuracy of the final threshold estimate or significantly alter the final threshold estimate obtained after the

completion of 50 trials. Additionally, increasing the slope parameter used in the QUEST procedure so that it was closer to the value found in the 4AFC used in this study may increase the efficiency and provide a more accurate threshold estimate. The information that this extra data would provide would enable a more detailed analysis of the QUEST procedure, and would indicate whether it was suitable for clinical use, or was more suited to an experimental setting where time constraints are less of an issue allowing a larger number of trials to be carried out.

The slope of psychometric functions estimated from multiple adaptive procedures targeting multiple thresholds could also be investigated, to determine whether accurate values could be obtained. Although it was beyond the scope of the current, study, the data obtained from the staircase tests (the SRT value) and the maximum-likelihood test (the 82% estimate) could be used to determine slope values, which could then be compared to the slope of the performance-intensity curves obtained from the conventional speech audiometry data. Because the reliability of the QUEST estimates was poor, the technique for determining a higher point on the psychometric curves would obviously require modifications before the procedure could be used to construct accurate slope estimates.

Future research should also address the limitations of this study, so that the adaptive speech tests follow the clinical protocols for speech testing more closely, and are applicable to a wider population. Additionally, more emphasis should be placed on determining what constitutes the ‘efficiency’ of conventional speech audiometry, so that valid comparisons can be made with regard to the efficiency of adaptive procedures versus methods of constant stimuli.

6.8 Conclusions

The current study sought to determine the optimal adaptive procedures for use in clinical speech tests and compare the administration time, accuracy, efficiency and reliability of these tests with the current form of speech audiometry as it is carried out in Audiology clinics. The results revealed the inadequacies of conventional speech audiometry, and the potential advantages of using adaptive staircase speech tests in their place.

In particular, this study provided evidence that adaptive staircase procedures are the most suitable procedures for testing speech discrimination abilities in a clinical Audiology setting. Both the closed-set and open-set forms of the adaptive staircase tests showed specific advantages over the current form of speech audiometry that employs a method of constant stimuli, most notably their increased reliability and greater inter-test consistency. In terms of the step size variations, the procedures employing a larger step size preliminary phase tended to show time and efficiency advantages over the equivalent tests employing constant step sizes throughout, with no loss of accuracy or reliability. The optimal termination criteria employed with each test differed, with the ‘within +4/-1 dB of the final threshold estimate’ requiring the fewest trials when used in conjunction with the closed-set adaptive test, and the ‘within ± 2.5 dB of the final threshold estimate’ requiring the fewest trials when used in conjunction with the open-set adaptive test.

Both the closed-set and open-set adaptive tests showed specific advantages; however, these were predominantly due to the different response formats and the scoring systems they employed, as opposed to the adaptive procedures themselves. The closed-set adaptive tests showed time advantages over the open-set adaptive tests, with no requirements for a clinician to score responses as this was performed automatically by the UC MAST program. This makes these closed-set adaptive tests more suitable in situations where patients have articulation difficulties or where clinician-patient time is limited. In contrast, the open-set tests required more time to complete but tended to yield lower SRT results that were closer to the PTA. This implies that these open-set adaptive tests may be more suitable in situations where an accurate value for an individual’s best possible threshold is required, as the phonemic scoring system they employed is able to score partially correct responses, thereby fine-tuning the final threshold estimate to a greater degree than the closed-set adaptive tests that employed a whole-word scoring system.

Although the maximum-likelihood QUEST procedure did not perform well in this study with respect to accuracy and reliability, the parameters chosen for its implementation were not optimal for a 4AFC task and therefore there is scope to potentially improve this procedure in subsequent studies. Using a greater number of trials and tracking the most efficient

threshold would help determine whether the QUEST procedure is suitable for clinical use, or whether it is best used in experimental settings.

Overall, there is enormous potential for the use of adaptive staircase procedures (both closed-set and open-set) in the clinical Audiology setting for the determination of speech discrimination information. The good performance of the open-set speech tests used in this study, and the lack of any previous research concerning adaptive procedures coupled with open-set tasks, suggests that further research into this area could be promising. The implementation of adaptive procedures into open-set tasks in both the Audiological domain and other domains could open up a range of novel uses for adaptive procedures, and possibly widen their use in both the clinical and experimental settings.

7. Appendices

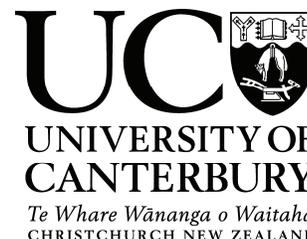
Appendix I: Project Information Sheets, Consent Form and Participant Questionnaire

Project Information Sheet: Part A

College of Science

Ms Brenna Sincock, Master of Audiology student
 Department of Communication Disorders
 Tel: +64 3 364 2987 ext. 7085, Fax: + 64 364 2760
 Email: bpl20@student.canterbury.ac.nz

Dr Greg O'Beirne, Lecturer in Audiology
 Department of Communication Disorders
 Tel: +64 3 364 2987 ext. 7085, Fax: +64 3 364 2760
 Email: gregory.obeirne@canterbury.ac.nz



PROJECT INFORMATION

A New Clinical Method for Testing Speech Hearing Ability - “Adaptive Speech Testing”

When testing a person’s hearing, it is important to test how well he/she can hear speech, because speech is what most people need to hear in order to communicate in their day-to-day lives. Speech testing is, therefore, an important part of the assessment performed in Audiology clinics. Currently in New Zealand Audiology clinics, speech testing is carried out using a method that involves patients listening to and repeating words presented in lists of fixed volumes. In comparison, a new method known as “adaptive speech testing” involves adjusting the volume of each word that is presented, according to the patient’s performance during the test. This makes the new method of adaptive speech testing more flexible than standard speech testing, because the volume of each word is tailored to the individual’s speech hearing ability. In addition, the new method of adaptive speech testing has the potential to provide accurate results in a shorter amount of time than standard speech testing, which is important for optimizing an Audiologist’s time with a patient.

The purpose of this Masters study is to determine if adaptive speech tests are suitable for use in Audiology clinics, and whether they provide any advantages over standard speech testing.

Your participation in this study is entirely voluntary (your choice). If you decide to participate, all testing will take place at the University of Canterbury Speech and Hearing Clinic. If you decide not to participate, this will in no way affect any further Audiological assessments or treatment that you may require.

The results of the tests you perform will be used in the current study, as well as possibly being used in future studies within the University of Canterbury. If you agree to participate, all identifying information, such as your name and address, will be kept confidential.

Part A: Creating the Adaptive Speech Test

The purpose of this part of the study is to determine the most suitable type of adaptive speech test (and the specific features of the test) to use in an Audiology clinic to test speech hearing ability. To determine the best type of adaptive speech test and the best combination of features, we need to test your performance on a range of different adaptive speech tests that use different features.

You will be required to attend two testing sessions, which each last approximately one hour. At the beginning of the first testing session you will be asked to fill out a short questionnaire to determine your suitability for participation in the study, and your hearing will be assessed if this has not already been done. The tasks that you will be required to perform in each session will include number of adaptive speech tests. These will involve listening to words and either repeating the word, or using a mouse to select the word from a group of alternative answers on a computer screen.

This project has been given ethical approval by the Upper South A Regional Ethics Committee and the University of Canterbury Humans Ethics Committee. The tests are perfectly safe and will in no way cause you any discomfort or harm. Nonetheless, you may end the tests at any time and are free to discontinue participation in this study, including withdrawal of any information you have provided.

If you have any queries or concerns regarding your rights as a participant in this study, you may wish to contact an independent Health and Disability Advocate, as follows:

South Island 0800 377 766

Free Fax (NZ wide) 0800 2787 7678 (0800 2 SUPPORT)

Email (NZ wide) advocacy@hdc.org.nz

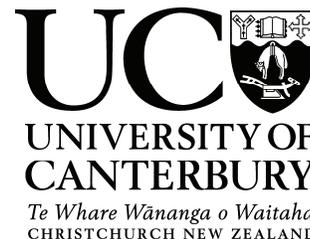
Thank you for choosing to take part in this study. Your participation is greatly appreciated.

Project Information Sheet: Part B

College of Science

Ms Brenna Sincock, Master of Audiology student
 Department of Communication Disorders
 Tel: +64 3 364 2987 ext. 7085, Fax: + 64 364 2760
 Email: bpl20@student.canterbury.ac.nz

Dr Greg O'Beirne, Lecturer in Audiology
 Department of Communication Disorders
 Tel: +64 3 364 2987 ext. 7085, Fax: +64 3 364 2760
 Email: gregory.obeirne@canterbury.ac.nz



PROJECT INFORMATION

A New Clinical Method for Testing Speech Hearing Ability - “Adaptive Speech Testing”

When testing a person’s hearing, it is important to test how well he/she can hear speech, because speech is what most people need to hear in order to communicate in their day-to-day lives. Speech testing is, therefore, an important part of the assessment performed in Audiology clinics. Currently in New Zealand Audiology clinics, speech testing is carried out using a method that involves patients listening to and repeating words presented in lists of fixed volumes. In comparison, a new method known as “adaptive speech testing” involves adjusting the volume of each word that is presented, according to the patient’s performance during the test. This makes the new method of adaptive speech testing more flexible than standard speech testing, because the volume of each word is tailored to the individual’s speech hearing ability. In addition, the new method of adaptive speech testing has the potential to provide accurate results in a shorter amount of time than standard speech testing, which is important for optimizing an Audiologist’s time with a patient.

The purpose of this Masters study is to determine if adaptive speech tests are suitable for use in Audiology clinics, and whether they provide any advantages over standard speech testing.

Your participation in this study is entirely voluntary (your choice). If you decide to participate, all testing will take place at the University of Canterbury Speech and Hearing Clinic. If you decide not to participate, this will in no way affect any further Audiological assessments or treatment that you may require.

The results of the tests you perform will be used in the current study, as well as possibly being used in future studies within the University of Canterbury. If you agree to participate, all identifying information, such as your name and address, will be kept confidential.

Part B: Clinical Testing Phase

You will be required to attend one testing session, which will last approximately forty-five minutes. If you are a client of the University of Canterbury Speech and Hearing Clinic, we will attempt to have this testing session coincide with your Audiology/hearing aid appointment. At the beginning of the testing session you will be asked to fill out a short questionnaire to determine your suitability for participation in the study, and you will be given a brief hearing screen.

During the main testing phase, you will be required to perform five speech tests, which will involve listening to words and either repeating the word or using a mouse to select the word from a group of alternatives on a computer screen.

This project has been given ethical approval by the Upper South A Regional Ethics Committee and the University of Canterbury Human Ethics Committee. The tests are perfectly safe and will in no way cause you any discomfort or harm. Nonetheless, you may end the tests at any time and are free to discontinue participation in this study, including withdrawal of any information you have provided.

If you have any queries or concerns regarding your rights as a participant in this study, you may wish to contact an independent Health and Disability Advocate, as follows:

South Island 0800 377 766

Free Fax (NZ wide) 0800 2787 7678 (0800 2 SUPPORT)

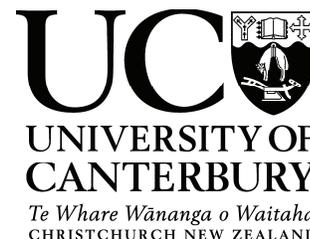
Email (NZ wide) advocacy@hdc.org.nz

Thank you for choosing to take part in this study. Your participation is greatly appreciated.

Consent Form**College of Science**

Ms Brenna Sincock, Master of Audiology student
 Department of Communication Disorders
 Tel: +64 3 364 2987 ext. 7085, Fax: + 64 364 2760
 Email: bpl20@student.canterbury.ac.nz

Dr Greg O'Beirne, Lecturer in Audiology
 Department of Communication Disorders
 Tel: +64 3 364 2987 ext. 7085, Fax: +64 3 364 2760
 Email: gregory.obeirne@canterbury.ac.nz

**CONSENT FORM****A New Clinical Method for Testing Speech Hearing Ability - "Adaptive Speech Testing"****Part B: Clinical Testing Phase****DECLARATION**

I (the participant) have read and understood the description and requirements of this study, as outlined in the attached information sheet dated 8 June 2007. I have had the opportunity to ask questions about the study, and am satisfied with the answers I have been given. On this basis, I agree to participate in this study, realising that participation is voluntary (my choice), and that I may withdraw at any time without prejudice.

I understand that my participation in this study is confidential and that no material which could identify me will be used in any reports on this study.

I agree that research data gathered in this study may be published and used in future studies. I provide consent for this publication and the reuse of the data with the understanding that my name and any other identifying information will not be used.

I wish to receive a copy of the results:

YES

NO

Participant: _____

Date _____

Project Explained by: _____

Date _____

Participant Questionnaire

Participant Identification Number: _____ Date: _____

**CLINICAL APPLICABILITY OF ADAPTIVE SPEECH TESTING: A COMPARISON
OF THE TIME EFFICIENCY, ACCURACY AND RELIABILITY OF ADAPTIVE
SPEECH TESTS WITH CONVENTIONAL SPEECH AUDIOMETRY****Participant Questionnaire**

Age: _____

	YES	NO
Do you have any literacy problems?	<input type="checkbox"/>	<input type="checkbox"/>
Do you have any speech problems?	<input type="checkbox"/>	<input type="checkbox"/>
Are you fluent in the English language?	<input type="checkbox"/>	<input type="checkbox"/>
Can you comfortably see the test word on the computer screen in front of you?	<input type="checkbox"/>	<input type="checkbox"/>

Appendix II: Instructions to Participants

Non-Adaptive Open-Set Test

In this speech test you will hear a man's voice telling you to say different words. He will say things such as "Say dog" or "Say cat." What I want you to do is repeat back the word he tells you to say – for example if you heard the words "Say dog" you would say "dog." The words will be presented in lists of ten. The first list will be at conversational level, the second list will be at a softer level, and the third list at a very soft level. If you do not hear a word it is important to have a guess or say part of the word, because you do get part marks if you get part of the word right. Try to be as fast and as accurate as possible.

Do you have any questions about what you have to do for this test?

Adaptive Open-Set Test

In this speech test you will hear a man's voice telling you to say different words. He will say things such as "Say dog" or "Say cat". What I want you to do is repeat back the word he tells you to say – for example if you heard the words "Say dog" you would say "dog." Each word will be presented at a different level that will depend on how much of the previous word you got correct. There will be a total of 40 words presented. If you do not hear a word it is important to have a guess or say part of the word, because you do get part marks if you get part of the word right. Try to be as fast and as accurate as possible.

Do you have any questions about what you have to do for this test?

Adaptive Closed-Set Tests

In this speech test you will hear a man's voice saying different words. After each word, four different words will appear on the computer screen in front of you. What I want you to do is use the mouse to click on the word that you think you heard. There will be a total of 40-50 words presented. If you are unsure or did not hear the word, it is important to have a

guess, as the test cannot continue until you have chosen an answer. Try to be as fast and as accurate as possible.

Do you have any questions about what to do for this test?

Non-Adaptive Closed-Set Test

You will hear a man's voice saying different words. After each word, I will show you four different words in a booklet. What I want you to do is point to the word that you think you heard. The words will be presented in two lists of twenty. The first list will be at a conversational level, while the second will be at a softer level. If you are unsure or did not hear the word, it is important to have a guess. Try to be as fast and as accurate as possible.

Do you have any questions about what to do for this test?

Appendix III: Example of a Complete Test Data Set (for Participant Seven in the Preliminary Testing Phase)

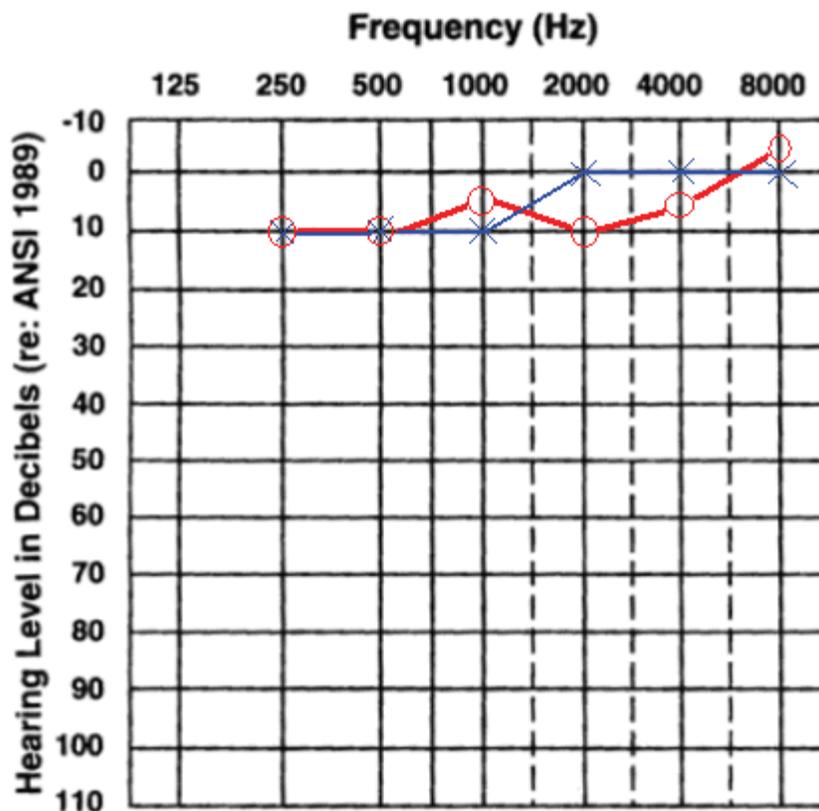


Figure 34. Audiogram showing hearing thresholds for Participant Seven (PTA for the right ear = 8.3 dB HL). Circles represent thresholds for the right ear; crosses represent thresholds for the left ear.

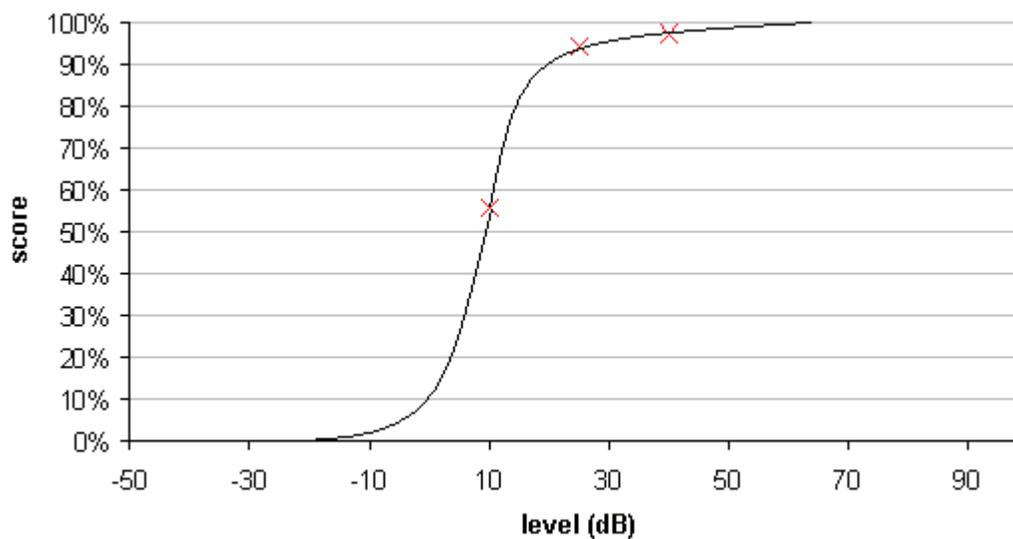


Figure 35. Speech curve for Participant Seven, estimated from the results of conventional speech audiometry (Speech Test Data: 97% at 40 dB, 94% at 25 dB, and 56% at 10 dB; SRT estimate = 7.9 dB HL).

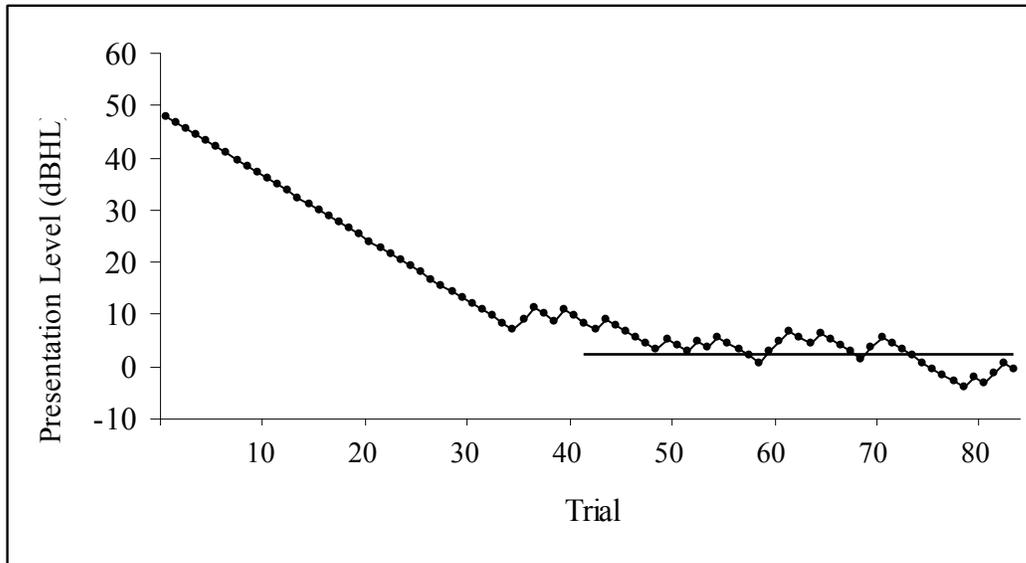


Figure 36. Closed-set constant step size tracking results for Participant Seven (Estimated SRT = 3.7 dB HL).

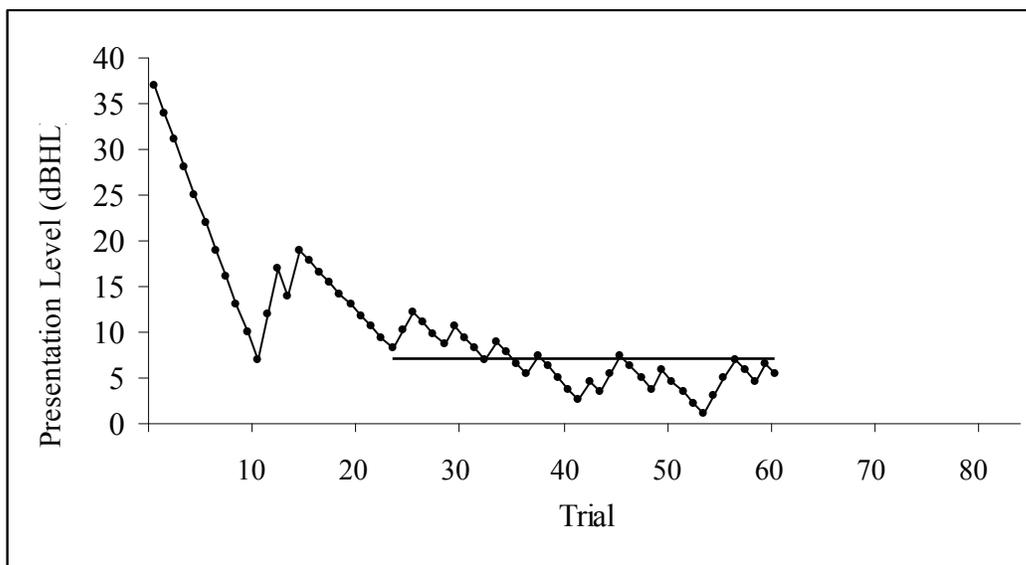


Figure 37. Closed-set larger step sizes at the beginning tracking results for Participant Seven (Estimated SRT = 7.2 dB HL).

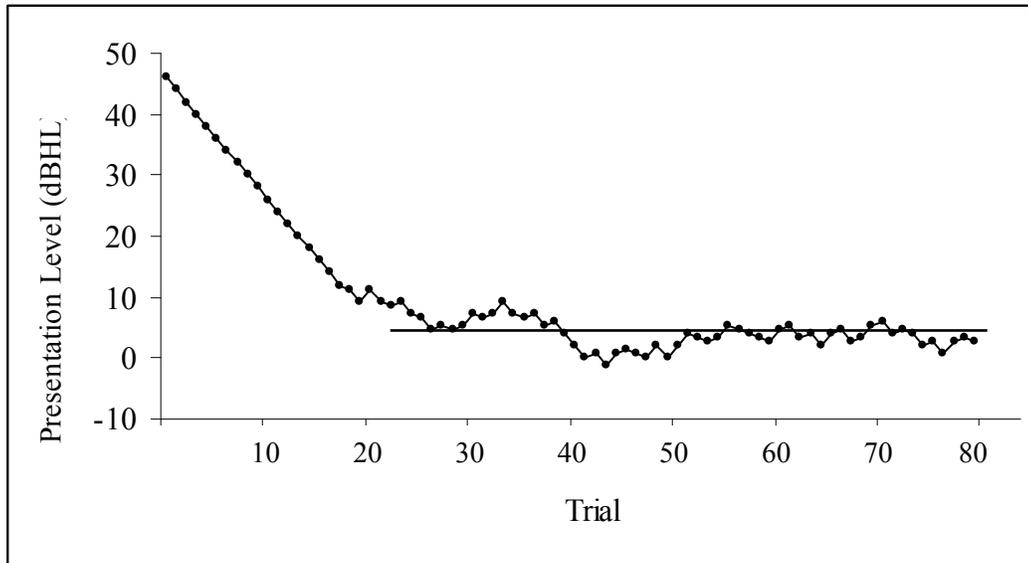


Figure 38. Open-set constant step size tracking results for Participant Seven (Estimated SRT = 3.4 dB HL).

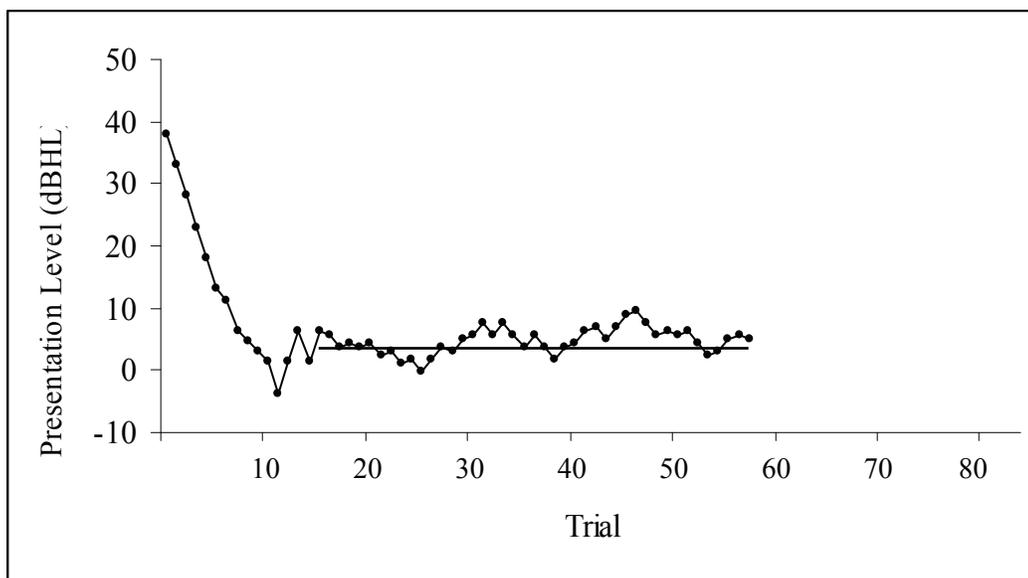


Figure 39. Open-set larger step sizes at the beginning tracking results for Participant Seven (Estimated SRT = 4.8 dB HL).

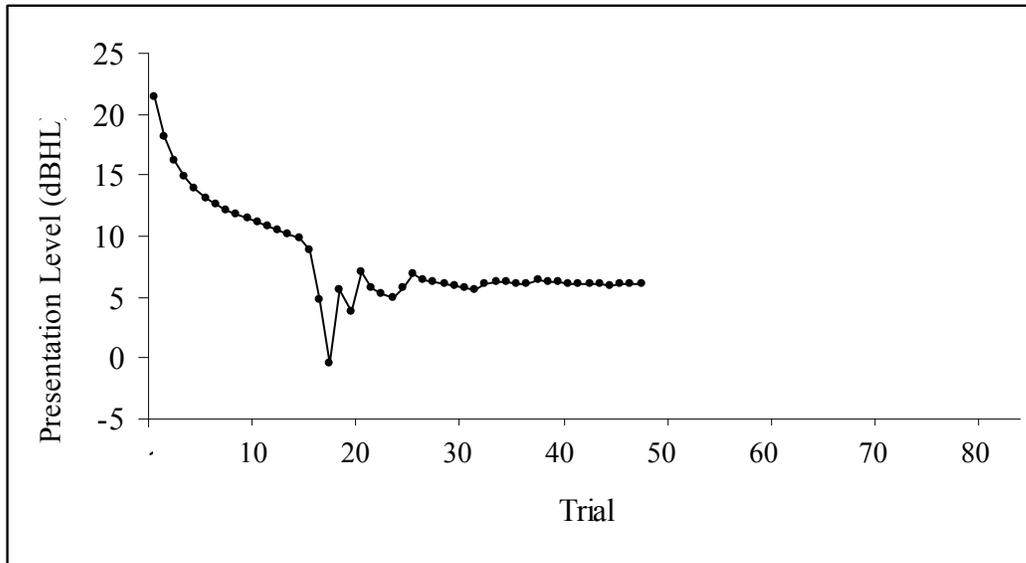


Figure 40. QUEST tracking procedure for Participant Seven (Estimated 82% threshold = 6.0 dB HL).

References

8. References

- American Speech-Language-Hearing Association Committee on Audiologic Evaluation. (1988). Guidelines for determining threshold level for speech. *ASHA*, 30, 85-89.
- Alcalá-Quintana, R., & García-Pérez, M. A. (2005). Stopping rules in Bayesian adaptive threshold estimation. *Spatial Vision*, 18(3), 347-374.
- Amitay, S., Irwin, A., Hawkey, D. J. C., Cowan, J. A., & Moore, D. R. (2006). A comparison of adaptive procedures for rapid and reliable threshold assessment and training in naive listeners. *The Journal of the Acoustical Society of America*, 119(3), 1616-1625.
- Amos, N. E., & Humes, L. E. (2007). Contribution of high frequencies to speech recognition in quiet and noise in listeners with varying degrees of high-frequency sensorineural hearing loss. *Journal of Speech, Language, and Hearing Research*, 50(4), 819-834.
- Anderson, A. J. (2003). Utility of a dynamic termination criterion in the ZEST adaptive threshold method. *Vision Research*, 43, 165-170.
- Bernstein, R. S., & Gravel, J. S. (1990). A method for determining hearing sensitivity in infants: the interweaving staircase procedure (ISP). *Journal of the American Academy of Audiology*, 1, 138-145.
- Boothroyd, A., & Nitttrouer, S. (1988). Mathematical treatment of context effects in phoneme and word recognition. *The Journal of the Acoustical Society of America*, 84(1), 101-114.
- Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision*, 10, 433-436.
- Brandy, W. T. (2002). Speech Audiometry. In J. Katz (Ed.), *Handbook of clinical Audiology* (5th ed., pp. 96-110). Baltimore: Lippincott, Williams and Wilkins.
- Bruce, D. J. (1956). Effects of context upon intelligibility of heard speech. In C. Cherry (Ed.), *Information Theory: Third London Symposium*. London: Butterworth.
- Buss, E., Hall, J. W., Grose, J. H., & Dev, M. B. (2001). A comparison of threshold estimation methods in children 6-11 years of age. *The Journal of the Acoustical Society of America*, 109(2), 727-731.
- Carhart, R. (1946). Monitored live voice as a test of auditory acuity. *Journal of the Acoustical Society of America*, 17, 339-349.
- Carhart, R. (1951). Basic principals of speech audiometry. *Acta Otolaryngology*, 40, 62-71.
- Carhart, R. (1971). Observations on relations between threshold for pure tones and for speech. *Journal of Speech and Hearing Disorders*, 36, 476-483.

- Carhart, R., & Jerger, J. (1959). Preferred method for clinical determination of pure-tone thresholds. *Journal of Speech and Hearing Disorders, 16*, 340-345.
- Carhart, R. B., & Porter, L. S. (1971). Audiometric configuration and prediction of threshold for spondees. *Journal of Speech and Hearing Research, 14*, 486-495.
- Clark, B., & Stewart, J. D. (1968). Comparison of three methods to determine thresholds for perception of angular acceleration. *The American Journal of Psychology, 81*(2), 207-216.
- Craig, C. H., Kim, B. W., Pecyna Rhyner, P. M., & Bowen Chirillo, T. K. (1993). Effects of word predictability, child development, and aging on time-gated speech recognition performance. *Journal of Speech and Hearing Research, 36*, 832-841.
- Crandell, C. C. (1991). Individual differences in speech recognition ability. *Ear and Hearing, 12*, Suppl:100S-108S.
- Dixon, W. J., & Mood, A. M. (1948). A method for obtaining and analyzing sensitivity data. *Journal of the American Statistics Association, 43*, 109-126.
- Eimas, P. D., Tajchman, G., Nygaard, L. C., & Marcus, D. J. (1996). Phonemic restoration and integration during dichotic listening. *Journal of the Acoustical Society of America, 99*(1141-1147).
- Falmagne, J. C. (1986). Psychophysical measurement and theory. In K. R. Boff, L. Kaufman & J. P. Thomas (Eds.), *Handbook of perception and human performance*. New York: John Wiley & Sons.
- Fletcher, H. (1950). A method for calculating hearing loss for speech from an audiogram. *Acta Otolaryngology, 90*, 26-37.
- Florentine, M., Marvit, P., & Buus, S. (2001). Maximum-likelihood yes-no procedure for gap detection: effect of track length. *Journal of the American Academy of Audiology, 12*, 113-120.
- Formby, C., Sherlock, L. P., & Green, D. M. (1996). Evaluation of a maximum likelihood procedure for measuring pure-tone thresholds under computer control. *Journal of the American Academy of Audiology, 7*, 125-129.
- García-Pérez, M. A. (1998). Forced-choice staircases with fixed step sizes: asymptotic and small-sample properties. *Vision Research, 38*, 1861-1881.
- García-Pérez, M. A., & Alcalá-Quintana, R. (2005). Sampling plans for fitting the psychometric function. *The Spanish Journal of Psychology, 8*(2), 256-289.
- Gelfand, S. A. (1998). Optimizing the reliability of speech recognition scores. *Journal of Speech, Language, and Hearing Research, 41*(5), 1088-1102.

- Gelfand, S. A. (2001). *Essentials in Audiology* (2nd ed.). New York: Thieme.
- Gelfand, S. A., & Silman, S. (1985). Functional hearing loss and its relationship to resolved hearing levels. *Ear and Hearing*, 6, 151-158.
- Goldstein, B. A., & Newman, C. W. (1994). Clinical masking: a decision-making process. In J. Katz (Ed.), *Handbook of Clinical Audiology* (4th ed., pp. 109-113). Baltimore: Williams & Wilkins.
- Green, D. M. (1990). Stimulus selection in adaptive psychophysical procedures. *Journal of the Acoustical Society of America*, 87(6), 2662-2674.
- Green, D. M. (1992). A maximum-likelihood method for estimating thresholds in a yes-no task. *Journal of the Acoustical Society of America*, 93(4), 2096-2105.
- Green, D. M. (1993). A maximum-likelihood method for estimating thresholds in a yes--no task. *The Journal of the Acoustical Society of America*, 93(4), 2096-2105.
- Green, D. M., Richards, V. M., & Forrest, T. G. (1989). Stimulus step size and heterogeneous stimulus conditions in adaptive psychophysics. *Journal of the Acoustical Society of America*, 88, 629-636.
- Gu, X., & Green, D. M. (1994). Further studies of a maximum-likelihood yes--no procedure. *The Journal of the Acoustical Society of America*, 96(1), 93-101.
- Hall, J. L. (1981). Hybrid adaptive procedure for estimation of psychometric functions. *Journal of the Acoustical Society of America*, 73, 663-667.
- Hall, J. L. (1983). A procedure for detecting variability of psychophysical thresholds. *Journal of the Acoustical Society of America*, 73(2), 663-667.
- He, N.-j., Dubno, J. R., & Mills, J. H. (1998). Frequency and intensity discrimination measured in a maximum-likelihood procedure from young and aged normal-hearing subjects. *The Journal of the Acoustical Society of America*, 103(1), 553-565.
- Heidenrich, S. M., & Turano, K. A. (1996). Speed discrimination under stabilized and normal viewing conditions. *Vision Research*, 36, 1819-1825.
- Helms, J., Weichbold, V., Baumann, U., von Specht, H., Schön, F., Müller, J., et al. (2004). Analysis of ceiling effects occurring with speech recognition tests in adult cochlear-implanted patients. *ORL: Journal for Oto-Rhino-Laryngology and Its Related Specialities*, 66(3), 130-135.
- Hughson, W., & Westlake, H. D. (1944). Manual for program outline for rehabilitation of aural casualties both military and civilian. *Transactions of the American Academy of Ophthalmology and Otolaryngology*, 48(Suppl.), 1-15.

- Johnson, C. A., Balwantray, C., & Shapiro, L. R. (1992). Properties of staircase procedures for estimating thresholds in automated perimetry. *Investigative Ophthalmology & Visual Science*, 33(10), 2966-2974.
- Kaernbach, C. (1991). Simple adaptive testing with a weighted up-down method. *Perception & Psychophysics*, 49, 227-229.
- Kaernbach, C. (2001). Adaptive threshold estimation with unforced-choice tasks. *Perception & Psychophysics*, 63(8), 1377-1388.
- Katz, D. R., & Elliot, L. L. (1978). *Development of a new children's speech discrimination test*. Paper presented at the annual meeting of the American Speech-Language-Hearing Association convention, Chicago.
- King-Smith, P. E. (1984). Efficient threshold estimates from yes-no procedures using few (about 10) trials. *American Journal of Optometry & Physiological Optics*, 61(119P).
- King-Smith, P. E., Grigsby, S. S., Vingrys, A. J., Benes, S. C., & Supowit, A. (1994). Efficient and un-biased modifications of the QUEST threshold method: Theory, simulations, experimental evaluation, and practical implementation. *Vision Research*, 34, 885-912.
- Kollmeier, B., Gilkey, R. H., & Sieben, U. K. (1988). Adaptive staircase techniques in psychoacoustics: A comparison of human data and a mathematical model. *The Journal of the Acoustical Society of America*, 83(5), 1852-1862.
- Kontsevich, L. L., & Tyler, C. W. (1999). Bayesian adaptive estimation of psychometric slope and threshold. *Vision Research*, 39, 2729-2737.
- Lam, C. F., Dubno, J. R., & Mills, J. H. (1999). Determination of optimal data placement for psychometric function estimation: A computer simulation. *Journal of the Acoustical Society of America*, 106(4), 1969-1976.
- Leek, M. R. (2001). Adaptive procedures in psychophysical research. *Perception & Psychophysics*, 63(8), 1279-1292.
- Leek, M. R., Hanna, T. E., & Marshall, L. (1992). Estimation of psychometric functions from adaptive tracking procedures. *Perception & Psychophysics*, 51(3), 247-256.
- Lerman, J. W., Ross, M., & McLaughlin, R. M. (1965). A picture-identification test for hearing-impaired children. *Journal of Audiological Research*, 5, 273-278.
- Levitt, H. (1971). Transformed Up-Down Methods in Psychoacoustics. *The Journal of the Acoustical Society of America*, 49(2B), 467-477.

- Linschoten, M. R., Harvey, L. O., Eller, P. M., & Jafek, B. W. (2001). Fast and accurate measurement of taste and smell thresholds using a maximum-likelihood adaptive staircase procedure. *Perception & Psychophysics*, *63*(8), 1330-1347.
- Mackie, K., & Dermody, P. (1986). Use of a monosyllabic adaptive speech test (mast) with young children. *Journal of Speech and Hearing Research*, *29*(2), 275-281.
- Marcell, M. M. (1995). Relationships between hearing and auditory cognition in Down's syndrome youth. *The Down Syndrome Educational Trust: Down Syndrome Research and Practice*, *3*(3), 75-91.
- Martin, F. N., Champlin, C. A., & Chambers, J. A. (1998). Seventh survey of audiometric practices in the United States. *Journal of the American Academy of Audiology*, *9*, 95-104.
- Marvit, P., Florentine, M., & Buus, S. (2003). A comparison of psychophysical procedures for level-discrimination thresholds. *The Journal of the Acoustical Society of America*, *113*(6), 3348-3361.
- McGaffin, A. J. (2007). *Development of a monosyllabic adaptive speech test for the identification of central auditory processing disorder*. Thesis. University of Canterbury
- McKee, S. P., Klein, S. A., & Teller, D. Y. (1985). Statistical properties of forced-choice psychometric functions: implications of probit analysis. *Perception & Psychophysics*, *37*(4), 286-298.
- Mitchell, M. L., & Jolley, J. M. (2004). *Research design explained*. Belmont: Wadsworth/Thomson Learning.
- Nagy, A. L., & Kamholz, D. W. (1995). Luminance discrimination, color contrast, and multiple mechanisms. *Vision Research*, *35*, 2147-2155.
- Nicholas, J. J., Heywood, C. A., & Cowey, A. (1996). Contrast sensitivity in one-eyed subjects. *Vision Research*, *36*, 175-180.
- Nilsson, M., Soli, S. D., & Sullivan, J. A. (1994). Development of the Hearing in Noise Test for the measurement of speech reception thresholds in quiet and in noise. *Journal of the Acoustical Society of America*, *95*(2), 1085-1099.
- Olsen, W. O., & Matkin, N. D. (1991). Speech Audiometry. In W. F. Rintelmann (Ed.), *Hearing Assessment* (2nd ed., pp. 39-140). Austin: Pro-Ed.
- Olsen, W. O., Van Tassel, D. J., & Speaks, C. E. (1997). Phoneme and word recognition for words in isolation and in sentences. *Ear and Hearing*, *18*(3), 175-188.

- Owens, E., & Schubert, E. D. (1977). Development of the California consonant test. *Journal of Speech and Hearing Research, 20*, 463-474.
- Pederson, O. T., & Studebaker, G. A. (1972). A new minimal contrasts closed-response-set speech test. *Journal of Audiological Research, 12*, 187-195.
- Pelli, D. G. (1997). The Video Toolbox software for visual psychophysics. Transforming numbers into movies. *Spatial Vision, 10*, 437-442.
- Pelli, D. G. (2005). Re: [psychtoolbox] Quest to manipulate mask intensity. Retrieved February 2, 2008, from <http://tech.groups.yahoo.com/group/psychtoolbox/message/3696>
- Pelli, D. G. (2006). Re: [psychtoolbox] Using QUEST for orientation discrimination task. Retrieved February 2, 2008, from <http://tech.groups.yahoo.com/group/psychtoolbox/message/5030>
- Pentland, A. (1980). Maximum-likelihood estimation: The best PEST. *Perception & Psychophysics, 28*, 377-379.
- Picard, M., Banville, R., Barbarosie, T., & Manolache, M. (1999). Speech audiometry in noise-exposed workers: the SRT-PTA relationship revisited. *Audiology, 38*, 30-43.
- Robbins, A. M., Renshaw, J. J., Miyamoto, R. T., Osberger, M. J., & Pope, M. L. (1988). *Minimal Pairs Test*. Indianapolis: Indiana University School of Medicine.
- Ross, M., & Lerman, J. (1970). A picture identification test for hearing impaired children. *Journal of Speech and Hearing Research, 13*, 44-53.
- Saberi, K., & Green, D. M. (1997). Evaluation of maximum-likelihood estimators in nonintensive auditory psychophysics. *Perception & Psychophysics, 59*, 867-876.
- Samuel, A. G. (1987). Lexical uniqueness effects on phonemic restoration. *Journal of Memory and Language, 26*, 36-56.
- Schefrin, B. E., Shinomori, K., & Werner, J. S. (1995). Contributions of neural pathways to age-related losses in chromatic discrimination. *Journal of the Optical Society of the American Academy, 12*, 1233-1241.
- Schlauch, R. S., Arnce, K. D., Olson, L. M., Sanchez, S., & Doyle, T. N. (1996). Identification of pseudohypacusis using speech recognition thresholds. *Ear and Hearing, 17*(3), 229-236.
- Schlauch, R. S., & Rose, R. M. (1990). Two-, three-, and four-interval forced-choice staircase procedures: Estimator bias and efficiency. *The Journal of the Acoustical Society of America, 88*(2), 732-740.

- Shadlen, M. (2002). *Re: [psychtoolbox] Why is threshold 82% in QUEST?* Retrieved February 2, 2008, from <http://tech.groups.yahoo.com/group/psychtoolbox/message/1277>
- Shelton, B. R., Picardi, M. C., & Green, D. M. (1982). Comparison of three adaptive psychophysical procedures. *The Journal of the Acoustical Society of America*, *71*(6), 1527-1533.
- Shelton, B. R., & Scarrow, I. (1984). Two-alternative versus three-alternative procedures for threshold estimation. *Perception & Psychophysics*, *35*(385-392).
- Silman, S., & Silverman, C. A. (1991). *Auditory Diagnosis: Principals and Applications*. San Diego: Academic Press.
- Silman, S., & Silverman, C. A. (1997). *Auditory Diagnosis: Principals and Applications*. San Diego: Singular Publishing Group.
- Simpson, W. A. (1988). The method of constant stimuli is efficient. *Perception & Psychophysics*, *44*, 433-436.
- Smallman, H. S., & MacLeod, D. I. A. (1994). Size-disparity correlation in stereopsis at contrast threshold. *Journal of the Optical Society of the American Academy*, *11*, 2169-2183.
- Snowden, R. J., Hess, R. F., & Waugh, S. J. (1995). The processing of temporal modulation at different levels of retinal illuminance. *Vision Research*, *35*, 775-789.
- Stach, B. A., Davis-Thaxton, M. L., & Jerger, J. (1995). Improving the efficiency of speech audiometry: computer-based approach. *Journal of the American Academy of Audiology*, *6*, 330-333.
- Stephens, S. D. G. (1976). The input for a damaged cochlea: a brief review. *British Journal of Audiology*, *10*, 97-101.
- Swanson, W. H. (1996). S-cone spatial contrast sensitivity can be independent of pre-receptoral factors. *Vision Research*, *36*, 3549-3555.
- Taylor, M. M., & Creelman, C. D. (1967). PEST: Efficient Estimates on Probability Functions. *The Journal of the Acoustical Society of America*, *41*(4A), 782-787.
- Taylor, M. M., Forbes, S. M., & Creelman, C. D. (1983). PEST reduces bias in forced choice psychophysics. *The Journal of the Acoustical Society of America*, *74*(5), 1367-1374.
- Treutwein, B. (1995a). Adaptive psychophysical procedures. *Vision Research*, *35*, 2503-2522.

- Treutwein, B. (1995b). Adaptive psychophysical procedures. *Vision Research*, 35(17), 2503-2522.
- Treutwein, B. (1997). YAAP: Yet another adaptive procedure. *Spatial Vision*, 11, 129-134.
- Turano, K. A., & Heidenrich, S. M. (1996). Speed discrimination of distal stimuli during smooth pursuit eye motion. *Vision Research*, 36, 3507-3517.
- Turpin, A., McKendrick, A. M., Johnson, C. A., & Vingrys, A. J. (2002). Development of efficient threshold strategies for frequency doubling technology perimetry using computer simulation. *Investigative Ophthalmology & Visual Science*, 43(2), 322-331.
- Verdon, W., & Haegerstrom-Portnoy, G. (1996). Mechanisms underlying the detection of increments in parafoveal retina. *Vision Research*, 36, 373-390.
- Voltz, H., & Zanker, J. M. (1996). Hyperacuity for spatial localization of contrast-modulated patterns. *Vision Research*, 36, 1329-1339.
- Warren, R. M. (1970). Perceptual restoration of missing speech sounds. *Science*, 167, 392-393.
- Watson, A. B., & Fitzhugh, A. (1990). The method of constant stimuli is inefficient. *Perception & Psychophysics*, 47(1), 87-91.
- Watson, A. B., & Pelli, D. G. (1983). QUEST: A Bayesian adaptive psychometric method. *Perception & Psychophysics*, 33(2), 113-120.
- Weibull, W. (1951). A statistical distribution function of wide applicability. *Journal of Applied Mechanics - Transactions of the American Society of Mechanical Engineers*, 18(3), 293-297.
- Wetherill, G. B., & Levitt, H. (1965). Sequential estimation of points on a psychometric function. *British Journal of Mathematical and Statistical Psychology*, 18, 1-10.
- Wolfson, S. S., & Graham, N. (2000). Exploring the dynamics of light adaptation: The effects of varying the flickering backgrounds duration in the probed-sinewave paradigm. *Vision Research*, 40, 2277-2289.
- Zanker, J. M., & Hüggen, I. S. (1994). Interaction between primary and secondary mechanisms in human motion perception. *Vision Research*, 34, 1255-1266.
- Zera, J. (2004). Speech intelligibility measured by adaptive maximum-likelihood procedure. *Speech Communication*, 42, 313-328.
- Zwislocki, J., Maire, F., Feldman, A. S., & Rubin, A. (1958). On the effect of practice and motivation on the threshold of audibility. *Journal of the Acoustical Society of America*, 30, 254-262.

